Workshop Report
Development of an Evidence-Based Risk Assessment Framework

Krewski, D.[1]*, Saunders-Hastings, P.[1], Baan, R.A.[2], Barton-Maclaren, T.S.[3], Browne, P.[4], Chiu, W.A.[5], Gwinn, M.[6], Hartung, T.[7], Kraft, A.[8], Lam, J.[9], R. Jeffery Lewis[10], Sanaa, M.[11], Morgan, R.L.[12], Paoli, G.[13], Rhomberg, L.[14], Rooney, A.[8], Sand, S.[15], Schünemann, H.J.[12], Straif, K.[2], Thayer, K.[6], Tsaioun, K.[17]

[1] School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada
[2] The *IARC Monographs Programme*, International Agency for Research on Cancer, Lyon, France (retired)
[3] Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Canada
[4] Organization for Economic Cooperation and Development, Paris, France
[5] College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, USA
[6] U.S. Environmental Protection Agency, Research Triangle Park, USA
[7] Center for Alternatives to Animal Testing, Johns Hopkins University, Baltimore, USA
[8] U.S. Environmental Protection Agency, Washington, USA
[9] Cal State East Bay, San Francisco, USA
[10] ExxonMobil Biomedical Sciences, Inc. Annandale, USA
[11] Agence Nationale Sécurité Sanitaire Alimentaire Nationale, Paris, France
[12] Faculty of Health Sciences, McMaster University, Hamilton, Canada
[13] Risk Sciences International, Ottawa, Canada
[14] Gradient, Cambridge, USA
[15] Swedish National Food Agency, Upsala, Sweden
[17] Evidence-Based Toxicology Collaboration, Johns Hopkins University, Baltimore, USA

*Corresponding author: Patrick.saundershastings@gmail.com

Summary

Assessment of potential human health risks associated with environmental and other agents requires careful evaluation of all available and relevant evidence for the agent of interest, including both data-rich and data-poor agents. With the advent of new approach methodologies in toxicological risk assessment, guidance on integrating evidence from multiple evidence streams is needed to ensure that all available data is given due consideration in both qualitative and quantitative risk assessment. The present report summarizes the discussions among academic, government, and private sector participants from North America and Europe in an international workshop convened to explore the development of an evidence-based risk assessment framework, taking into account all available evidence in an appropriate manner in order to arrive at the best possible characterization of potential human health risks, and associated uncertainty. Although consensus among workshop participants was not a specific goal, there was general agreement on the key considerations involved in evidence-based risk assessment incorporating 21st century science into human health risk assessment. These considerations have been embodied into an overarching prototype framework for evidence integration that will be explored in more depth in a follow-up meeting.

Keywords: risk assessment; environmental agents; population health; new approach methodologies; evidence integration.

1. Introduction and Background

Risk science has evolved into a well-established interdisciplinary practice incorporating diverse data and methods in order to characterize population health risks and inform decision-making. Risk science has benefitted from advances in biology and toxicology over the last decade, providing powerful new tools and technologies — including high-throughput in vitro screening and computational toxicology — that can be used to better assess risks to population health. Risk science has also benefitted from advances in molecular and genetic epidemiology which, combined with concomitant advances in exposure science, permit direct estimation of risk in human populations. These and other advances have been incorporated into a framework for the next generation of risk science proposed by Krewski et al. (2014), which was based on work completed under the US Environmental Protection Agency (EPA) NexGen program, with input from a large number of stakeholders from North America and Europe.

An important aspect of the evolution of risk science is the desire to ensure that risk decisions are based on the best available scientific evidence, with this evidence identified and evaluated in accordance with appropriate processes and criteria. This trend is consistent with the evolution of evidence-based medicine, which makes use of current best evidence in making clinical decisions about the care of individual patients (Masic, Miokovic, & Muhamedagic, 2008; Sackett, 1997). More recently, the concept of evidence-based toxicology has emerged under the leadership of investigators at the Johns Hopkins Bloomberg School of Public Health (Stephens et al., 2013). Like evidence-based medicine, evidence-based toxicology seeks to ensure that the best available data is used in toxicological risk assessment.

The present initiative seeks to build on the scientific advances covered above, and the trends towards evidence-based decision making in multiple disciplines, to derive a framework for evidence-based risk assessment that incorporates all relevant data needed to support risk decision-making in a transparent and objective manner. The specific objectives of this project are:

(1) to develop a framework for evidence-based risk assessment describing how all relevant evidence relating to a specific risk decision should be assembled and evaluated;
(2) to conduct case study prototypes to evaluate the utility of the framework and demonstrate its application in practice; and
(3) to lay out a knowledge translation action plan to support the adoption and use of the framework for evidence-based risk assessment in decision making.

2. Evidence for Causation

Establishing causality requires a careful evaluation of the available evidence for and against a causal association between exposure and outcome. Evaluating evidence for causality can be a complex undertaking, particularly in the presence of diverse sources of information which may report inconsistent findings and which may be of unequal relevance or reliability. A systematic review can be used to summarize the available evidence in a comprehensive and reproducible manner (Wang, Gomes, Cashman, Little, & Krewski, 2014). Although not all systematic reviews are designed to evaluate causality, there has been a trend towards including causality evaluation as a component of systematic review in recent years. Historically, the Hill criteria (strength, consistency, specificity, temporality, biological gradient, plausibility,

coherence, experiment, and analogy) have provided useful general guidance on weighing the evidence for causality (Lucas & McMichael, 2005), though they were originally designed with only observational (epidemiologic) data in mind, and do not address other aspects of causality determination, such as consideration of experimental data and integrating different sources of evidence. The grading of recommendations, assessment development, and evaluation (GRADE) approach incorporates aspects of the considerations for causality identified by Hill as well as other considerations, providing an approach to evaluate the certainty of the body of evidence across the following domains: risk of bias, inconsistency, indirectness, imprecision, publication bias, magnitude of effect, dose-response gradient, and opposing residual confounding (Schunemann et al., 2008). On the other hand, the International Agency for Research on Cancer (IARC), for example, has developed a detailed approach for identifying agents that can cause cancer in humans, based on a careful evaluation of the available human, animal and mechanistic data (IARC, 2019).

Rhomberg and colleagues (2013) recently reviewed 50 different frameworks that have been proposed in different contexts in the interests of developing a "transparent and defensible" methodology for evaluating the evidence for causation. This review identified four key phases for such assessments: (1) defining the causal question and developing criteria for study selection, (2) developing and applying criteria for review of individual studies, (3) evaluating and integrating evidence and (4) drawing conclusions based on inferences. Although a specific framework that would be widely applicable in different contexts was not proposed, this work serves to define important attributes of what a broadly applicable framework might include. Five years later, another review of the body of knowledge presented a framework with a similar four-step approach: (1) plan and scope the weight of evidence (WoE) assessment, (2) establish lines of evidence, (3) integrate line to assess WoE, and (4) summarize conclusions (Martin et al., 2018). While the specific principles, practices and approaches proposed by these two reviews may differ, together they offer a general approach to evaluating evidence for causation that can be refined and adapted as needed.

Several organizations have provided more detailed guidance for evaluating evidence of causation in various circumstances, depending in part on the nature of the available data (predominantly epidemiological or toxicological) and the risk decision context. Following a review of risk assessment approaches used by the US EPA's Integrated Risk Information System (IRIS), the US National Research Council (2014a) identified systematic review and evidence integration as key components of a qualitative and quantitative risk assessment paradigm for environmental chemicals. More broadly, the National Toxicology Program Office of Health Assessment and Translation developed a *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration.* In the context of establishing dietary reference intakes (DRIs) taking into account chronic disease outcomes, a committee of the National Academies of Sciences, Engineering, and Medicine ((The National Academies of Sciences Engineering and Medicine, 2017) developed *Guiding Principles for Developing Dietary Reference Intakes Based on Chronic Disease,* adopted GRADE (Guyatt et al., 2008) as the preferred approach to both establishing evidence of causation as well as for intake-response assessment.

3. Sources of Evidence

In conducting evidence-based risk assessment, it is important that all available and relevant sources of information be considered. Risk assessment may be informed by toxicological, epidemiological, clinical, surveillance, mechanistic and other data, all of which need to be considered collectively in order to ensure that the evidence base assembled to potentially support the assessment conclusions is appropriately comprehensive.

An important aspect of evidence-based risk assessment is the amount of data that may or may not be available to support the assessment. Cote and colleagues (2016) note that data-rich and data-poor risk decisions necessarily require different approaches and offer advice on what might be done in a data-poor situation where a risk decision must be made without the luxury of filling key data gaps.

In elaborating the proposed framework for evidence-based risk assessment, the strengths and limitations of different sources of information will be identified, and their complementary role in informing the overall assessment outlined. The framework will address both data-rich and data-poor risk decision contexts and establish minimum data requirements to support evidence-based risk decision making.

3.1 Current Approaches to Evidence Integration

3.1.1 EFSA

With a mandate to provide scientific expertise related to food and feed products in the European Union, the European Food Safety Authority (EFSA) is among the global leaders in hazard identification and risk assessment. Established in 2002 with funding from the European Union, EFSA has developed a variety of frameworks and approaches to support transparency, rigor and quality in evidence-based risk assessment for products under their remit (EFSA, 2010, 2014, 2015). While additional information on some of these is provided in a separate publication in this issue (Aiassa, Merten, & Martino, 2020), a brief summary of the 4-step framework for conducting a scientific risk assessment is included herein.

The 4-step process comprises “plan”, “do”, “verify” and “report” stages. Again, while additional information on the structure and application of this framework is provided in a separate publication (Aiassa et al., 2020), the approach can briefly be conceptualized as follows:

1. Plan: Formulate the key research question and (as relevant) associated sub-questions, outlining the methodology for answering the question(s) in a protocol developed *a priori*
2. Do: Execute the methodology outlined in the protocol to collect, analyze and leverage data to inform conclusions (specifics will depend on study design and type(s) of data being collected)
3. Verify: Compare the methodology taken with that outlined in the protocol, making note of any deviations from the original plan
4. Report: Promote transparency through the publication of relevant methodologies, assumptions, results and uncertainties

This approach has been piloted across EFSA with some success. For example, such an approach promotes the impartiality, rigour and overall scientific value of the assessment process — as well as the resulting

conclusions — by reducing the risk of bias from decisions made in light of the data collected (Munafo et al., 2017; Shamseer et al., 2015). Conversely, it was found that the implementation of the proposed approach among new or novice users was a resource-, effort- and time-intensive process, and EFSA continues to work towards building capacity and expertise to deliver scientific assessments that are both efficient and in line with current best practices in risk assessment. Nevertheless, EFSA's prioritization of the principles of impartiality, methodological rigour, transparency and public engagement point to the value of the promotion and integration of such a framework within everyday practice, recognizing that continued improvement should further advance EFSA's ability to deliver high-quality scientific assessments of relevance to public health promotion across the European Union.

3.1.2 EPA IRIS

Created in 1985 and located within the EPA Center for Public Health and Environmental Assessment, the IRIS Program conducts chemical hazard assessments (EPA, 2018). These assessments examine the health consequences of lifetime exposure to environmental chemicals and are both a primary source of certain chemical toxicity information used in support of regulatory and non-regulatory decisions within EPA program offices and regions, and important sources of information for other state, federal and international organizations.

The IRIS approach for assessment development (illustrated as interpreted by the NAS in **Figure 1**) has increasingly been framed through the lens of rigorous and transparent systematic review processes. Assessment development is part of a larger seven-step process for assessment review, which can be summarized as follows (NRC, 2011):

1. Complete draft IRIS assessment
2. Internal agency review
3. Science consultation on the draft assessment with other federal agencies and White House offices
4. Independent expert peer review, public review and comment, and public listening session
5. Revise assessment
6. A. Internal agency review and EPA clearance of final assessment
   B. EPA-led interagency science discussion
7. Post final assessment on IRIS

Assessments are intended to inform decisions related to hazard identification and dose-response assessments, which can be integrated with exposure assessments by EPA programs and regional offices to characterize potential public health risks associated with exposure to an environmental chemical or group of chemicals.

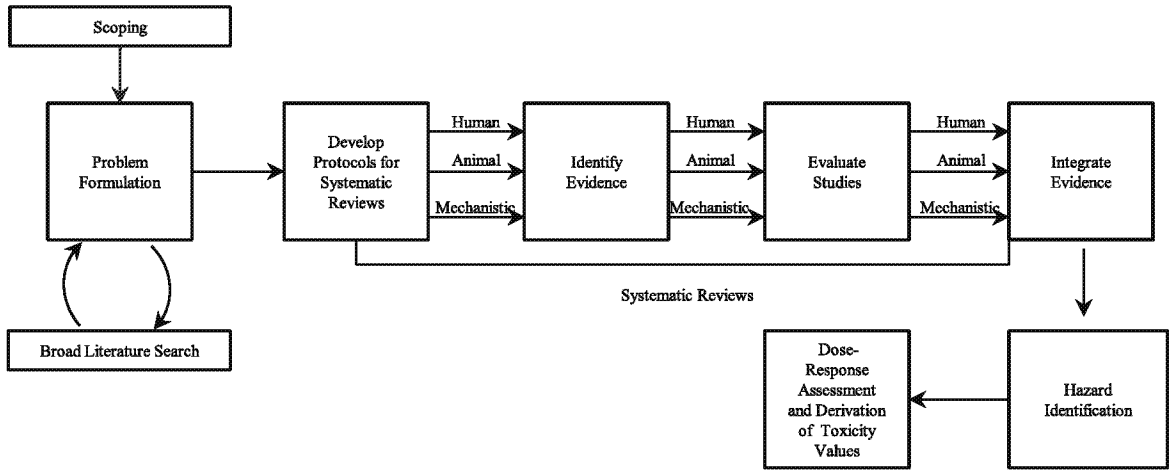


Figure 1. IRIS approach for assessment development, as interpreted by the NAS (NRC, 2014a).

IRIS supports the principle of transparency in their decision processes, with publicly available summaries and databases of chemical-specific evidence and assessment judgments provided since 1988 (EPA, 2018). Progress towards the application of best practices in systematic review and risk assessment has accelerated since 2011, when recognition of challenges in previous assessments motivated a commentary on the IRIS assessment development process by the NAS that was outside of the scope of the chemical-specific review (NRC, 2011). The resulting recommendations outlined a roadmap for a more systematic review process that triggered numerous planned enhancements to the assessment development process, including on-boarding of systematic review methodologies, adoption of the Health and Environmental Research Online (HERO) tool, and increased public engagement (EPA, 2018; National Academies of Sciences Engineering and Medicine, 2018).

In order to review changes and progress in the years following the 2011 NAS review, another committee was convened in 2014 (NRC, 2014a); systematic review and application of best practices in evidence integration were again identified as essential elements of environmental chemical human health assessment. Further improvements to the rigor of the IRIS process are being implemented, as reflected in a third assessment conducted in 2018 (National Academies of Sciences Engineering and Medicine, 2018), which concluded that the IRIS process — while still evolving to adapt to new scientific methodologies and data sources — had successfully undertaken reforms to improve the application and transparency of systematic review methodologies in chemical assessments. Moving forward, it was noted that new tools and approaches would be required to meet some of the outstanding recommendations from the 2014 assessment, "especially for incorporating mechanistic information and for integrating evidence across studies" (National Academies of Sciences Engineering and Medicine, 2018, p. 12).

3.1.3 Health Canada

The Canadian Environmental Protection Act (CEPA), 1999, serves as the main federal policy under which potentially hazardous environmental substances are assessed and regulated. Health Canada and Environment and Climate Change Canada work together to assess the potential for risk to the

environment and the general Canadian population associated with these substances and, as necessary, develop policies and risk management measures for their control. Since its ratification twenty years ago, over 23,000 environmental substances have been registered on the Canadian Domestic Substances List (DSL) (Krewski et al., 2019). The Canadian Chemicals Management Plan (CMP), launched in 2006 based on results of Categorization of the DSL and New Substances Notifications, further sought to evaluate the risk associated with 4,300 prioritized chemicals prior to 2020. Our knowledge of chemicals and emerging technologies continues to evolve. Therefore, moving forward it is important to continue to screen, integrate and consider new information and the increasing complexities of chemicals that may have the potential to cause harm to the environment or human health. Under the CMP, the identification of risk assessment priorities (IRAP)[1] approach is the ongoing prioritization activity for systematically collecting, consolidating and analyzing information for chemicals and polymers.

Given the ambitious timelines and number of chemicals to be assessed and addressed, an important element in Health Canada's success to date has been the development and application of the CMP Risk Assessment Toolbox (**Figure 2**). This Toolbox was developed to delineate the various types of approaches used to address the remaining substances or groups of substances prioritized under the CMP. To make best use of available information, gain efficiencies and ensure the ability to focus on substances of highest concern, the Risk Assessment Toolbox outlines three types of approaches that can be selected as appropriate and used in a fit-for-purpose manner based on the complexity of the assessment required (Health Canada, 2016):

- **Type 1 Approaches** use science-based policy responses such as referral of the assessment to a better-placed federal risk assessment program or documentation that the substance has already been addressed by an existing action or previous initiative under CEPA.
- **Type 2 Approaches** address substances using broad-based quantitative or qualitative approaches and apply conservative (protective) assumptions. Formal CEPA conclusions may or may not be made under section 64.
- **Type 3 Approaches** are applied for substances requiring a standard risk assessment approach including both hazard and exposure and may consider qualitatitive and quantitative lines of evidence in the determination of whether the substances or group of substances meet the criteria under section 64 of CEPA 1999. Further, the Toolbox proposes three approach subtypes spanning a continuum of complexity and methodology considerations in order to focus the risk assessment efforts.

Select examples of the types of approaches are noted in Figure 2 and the scientific details are described in the published CMP Science Approach Documents (SciADs). More information on the approaches, application and results, including the Threshold of Toxicological Concern, Ecological Risk Classification, and biomonitoring-based approaches can be found on the Chemical Substances webpage (Health Canada, 2020)[2].

[1] https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/identification-risk-assessment-priorities.html

[2] https://www.canada.ca/en/health-canada/services/chemical-substances/science-approach-documents.html

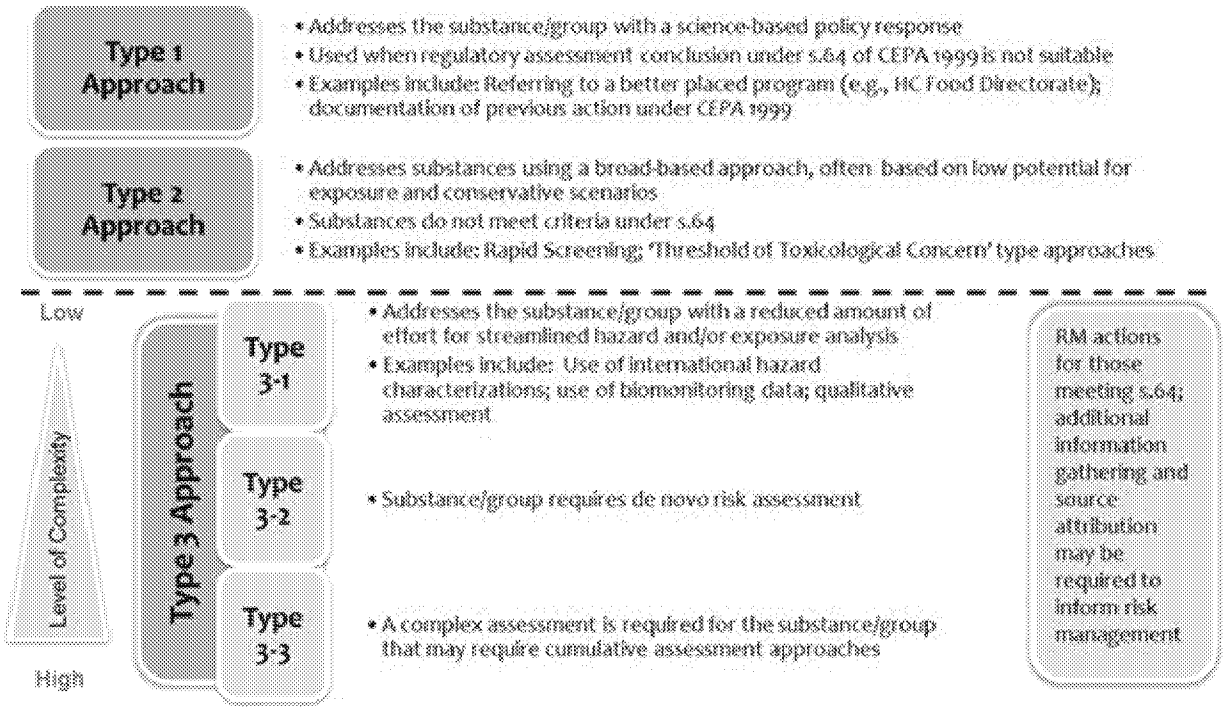


Figure 2. Chemical Management Plan Risk Assessment Toolbox (Health Canada, 2016).

As the Government of Canada embarks on the next phase of their chemicals assessment and management program, new approach methods (NAMs) are being considered and developed for inclusion in the CMP Risk Assessment Toolbox, particularly for Type 2 and 3 approaches, to rapidly and effectively identify the potential for risk in support of the 21st century paradigm shift in risk science. To date, there has been a high degree of success in advancing the use of new technologies and analytical tools through several case studies that have illustrated the practical and positive impacts of integrating multiple lines of evidence, including emerging science. A solid foundation and proof of concept has been illustrated for the application of several important NAM and computational toxicology including the example presented at the workshop on the use of Integrated Approaches for Testing and Assessment (IATA) for screening level risk assessment (Webster et al. 2019).

As Health Canada continues to advance chemicals assessment and management, NAM will be considered in the evolving risk assessment toolbox and incorporated into decision-making through the application of robust methodologies that are context specific and fit-for-purpose.

3.1.4 ANSES

Over the period 2015–2016, the French Agency for Food, Environmental and Occupational Health & Safety, ANSES (l'agence national de sécurité sanitaire alimentation, environment, travail) convened an expert panel to provide a critical review and advice on best practices in evaluation of the weight of evidence ("le poids des preuves") in the hazard identification step of the risk assessment process. ANSES is somewhat unique in this regard due to the breadth of their mandate. They sought to harmonize the application of weight-of-evidence concepts across multiple hazardous domains including public and

occupational exposures to chemical hazards, radiation hazards, nutrients and microbial hazards, but all within the hazard identification stage. In this way, questions such as "Does exposure to this chemical cause cancer?" and "Are these particular prions transmissible to humans?" would be answered in a rigorous and harmonized approach. The result of the process was a report with several findings and recommendations (Makowski et al., 2016).

Following the literature review, the panel described a four-step process that is similar in general structure to other frameworks in the literature (planning, evaluation of lines of evidence, integration of lines of evidence and reporting on the overall weight-of-evidence). The panel's recommendations to ANSES were described in line with this framework. Further work by ANSES has included, among other activities, consideration of the role of quantitative approaches to weight-of-evidence.

3.1.5 IARC

Several national and international health agencies have established programs with the aim of identifying agents and exposures that cause cancer in humans. The *IARC Monographs on the Identification of Carcinogenic Hazards to Humans* are published by the IARC and the World Health Organization (WHO). Each *IARC Monograph represents* the consensus of an international working group of expert scientists. The *Monographs* include a critical review of the pertinent peer-reviewed scientific literature as the basis for an evaluation of the weight of the evidence that an agent may be carcinogenic to humans. Published continuously since 1972, the scope of the *IARC Monographs* has expanded beyond chemicals to include complex mixtures, occupational exposures, lifestyle factors, physical and biologic agents, and other potentially carcinogenic exposures. To date, 120 *IARC Monograph* Volumes are available on-line.[3] and four more are in preparation. After the forthcoming publication of Volume 124, more than 1000 agents, mixtures, and exposures will have been evaluated. Among these, 120 have been characterized as carcinogenic to humans, 82 as probably carcinogenic to humans, and 311 as possibly carcinogenic to humans.

From the very beginning, there have been two criteria for consideration of an agent for evaluation: (a) there is evidence of human exposure and (b) there are published scientific data suggestive of carcinogenicity. For each agent considered, systematic reviews of the available scientific evidence on its carcinogenicity in humans and in experimental animals are conducted by an international working group of independent experts. Data on human exposure to the agent and toxicological data on pertinent mechanisms of carcinogenesis are also reviewed. An overall evaluation that integrates epidemiological and experimental cancer data as well as mechanistic evidence, most notably in exposed humans, is reached according to a structured process. Agents with 'sufficient evidence of carcinogenicity' in humans are assigned by default to the highest category, 'carcinogenic to humans' (IARC Group 1) whereas the categories of 'probably' (Group 2A) or 'possibly' (Group 2B) carcinogenic to humans, or 'not classifiable as to its carcinogenicity to humans' (Group 3) are assigned according to the combined strength of the human, animal and mechanistic evidence. Agents may be placed in a higher category when the evidence for a

[3] See https://monographs.iarc.fr/monographs-and-supplements-available-online/

relevant mechanism of carcinogenesis is sufficiently strong. The *IARC Monographs* classifications refer to the strength of the evidence for a cancer hazard, rather than to the level of cancer risk.

The *Preamble* to the *IARC Monographs* describes the objective and scope of the *Programme*, the scientific principles and procedures used in developing a *Monograph*, the types of evidence considered, and the scientific criteria that guide the evaluations. The *IARC Monographs* are prepared according to principles of scientific rigour, impartial evaluation, transparency, and consistency. The criteria defining those principles have evolved during the early years of the *Programme* and were outlined in the first *Preamble* (IARC, 1978), which has been refined and updated a dozen times since. In the recently revised *Preamble* (IARC, 2019) mechanistic evidence has been given a place that is equivalent to that of epidemiological evidence and evidence from animal studies. Mechanistic studies have gained in prominence, increasing in volume, diversity, and relevance to cancer hazard evaluation. The major change in the new *Preamble* is the introduction of systematic review of mechanistic data facilitated by the organization into Key Characteristics (Smith et al., 2016), which is now common practice since *Monograph* Volume 112. A useful overview of the major components of the most recent update of the Preamble is provided in a recent publication by Samet and colleagues (2019).

3.2 New Approach Methodologies

Following publication of the US National Research Council report, *Toxicity Testing in the 21st Century: A Vision and A Strategy* (NRC, 2007), there has been increasing emphasis on alternatives to animal testing in toxicological risk assessment. A mid-term update on progress made towards the realization of this vision over the original 20-year planning horizon has recently been prepared by Krewski and colleagues (2019). The broad suite of tools and strategies offering viable alternatives to animal testing is now referred to as new approach methodologies, or NAMs. The US Environmental Protection Agency (EPA, 2019) has developed a list of NAMs considered potentially relevant for evaluating chemical toxicity under the Toxic Substances Control Act. These new approaches can be broadly classified as including:

- computational toxicology and bioinformatics;
- high-throughput screening methods;
- testing of categories of chemical substances;
- tiered testing methods;
- in vitro studies;
- systems biology; and
- new or revised methods from validation bodies such as ICCVAM, ECCVAM, NICETAM, and OECD.[4].

The EPA list also includes a number of alternative test methods that have been validated by OECD, such as the *Bacterial Reverse Mutation Test* (OECD Test Guideline 471) and the *Performance-Based Test Guideline for Human Recombinant Estrogen Receptor (hrER) in vitro Assays* (OECD TG 493). The OECD

[4] ECCVAM: European Center for the Validation of Alternative Methods; ICCVAM: Interagency Coordinating Committee on the Validation of Alternative Methods; NICEATM: NTP Interagency Center for the Evaluation of Alternative Toxicological Methods; OECD: Organization for Economic Cooperation and Development.

Toolbox provides a rich suite of computational tools for evaluating quantitative structure-activity relationships (OECD, 2020).

The US EPA National Center for Computational Toxicology (NCCT) has recently developed a roadmap outlining an approach for making greater use of new approach methodologies (NAMs) (Thomas et al., 2019). Key elements of the EPA CompTox Blueprint include an emphasis on computational modeling and high-throughput approaches to supplement traditional approaches in chemical assessments for regulatory-decision making. On December 17, 2019, EPA hosted its First Annual Conference on the State of the Science on Development and Use of NAMs for Chemical Safety Testing. (EPA is currently preparing a summary of this meeting, that will be posted on their website when available.)

Andersen et al. (2019) have suggested a multi-level strategy for incorporating NAMs into toxicological risk assessment practice, seeking to deploy new methods in a context specific manner. Level 1 in the proposed strategy focuses on computational screening, with quantitative structure-activity relationships (QSAR)/read across, cheminformatics, and threshold of toxicological concern approaches used to assess bioactivity and high-throughput exposure modeling approaches used to evaluate potential human exposure. Level 2 relies on high-throughput in vitro screening to assess bioactivity through transcriptomics, high-content imaging and bioinformatics, along with judiciously chosen test batteries, to evaluate bioactivity. Refined exposure models may be used, along with high-throughput in vitro to in vivo extrapolation (HT-IVIVE), to estimate human doses. Level 3 invokes fit-for-purpose assays for bioactivity, including lower throughput cell-based assays and consideration of metabolism. More specific exposure models are employed, along with quantitative in vitro to in vivo extrapolation (q-IVIVE). Level 4 employs more complex in vitro assays for bioactivity, advanced systems such as organ chips, and tailored in vivo studies to confirm in vitro results. Tailored exposure models may also be employed at this level, including physiologically based pharmacokinetic (PBPK) models for in vivo species extrapolation.

As envisaged by Andersen et al. (2019), level 1 approaches may be most useful in the context of priority setting, with level 2 approaches more suited to screening level assessments. Level 3 approaches afford greater insight into toxicity pathways and are consistent with the vision put forward by the US National Research Council (NRC, 2007) for toxicity testing in the 21st century (TT21C). When required, level 4 approaches may provide additional data using more integrated assays at the biological system level. At each level, margins of exposure (MOEs) based on a comparison of predicted human doses to doses at which bioactivity is expected provide valuable information in support of context specific risk decisions.

As an example of a successful application of NAMs in risk assessment, OECD researchers have applied performance-based approaches to assess the reliability and accuracy of in vitro predictions (OECD 2019). Using estrogen receptor bioactivity models based on 18 HTS assays, 43 reference chemicals achieved a balanced accuracy of 86–95%; similar validation exercises for androgen receptor activity and anti-androgen activity also produced strong measures of validation performance. Encouraged by the success of these in vitro approaches, the OECD has published three guidance documents on incorporating NAMs into integrated assessment and testing approaches.

New approaches to risk decision making are increasingly being used by regulatory authorities worldwide, motivated in large part by the need for increased throughput in risk decision making. The US EPA used Attagene assays including multiple gene targets such as PPARα and NRF2 to rapidly evaluate the relative toxicity of eight dispersants in response to the Deepwater Horizon oil spill in the Gulf of Mexico in 2009, and this represents an early practical application of NAMS in emergency risk decision-making by the US EPA (Anastas, Sonich-Mullin, & Fried, 2010). With the diverse set of NAMs currently available, there are unprecedented opportunities to apply new high- and medium-throughput assays in support of human health risk assessment. The elaboration of strategies for deploying new approach methodologies in a systematic manner, such as that suggested by Andersen et al. (2019), will be of great value in choosing the most appropriate approaches to employ within specific risk decision contexts. While such efforts are both needed and welcome, the increased use of NAMs by regulatory authorities will also require new thinking on how to incorporate new types of data into weight of evidence evaluations.

4. Defining the Research Question

Formulating a clear and actionable research question creates structure in the approach to conducting systematic reviews and developing health guidance (Guyatt, Oxman, Kunz, Atkins, et al., 2011) Within the field of risk assessment, this question may be tailored for studies of exposure as a PECO question, which is used to outline the Population, Exposure, Comparator, and Outcomes (Morgan, Whaley, Thayer, & Schunemann, 2018). Guidance has been provided to help users operationalize the PECO question, as this informs many stages of the evidence review and quality assessment of the findings (**Section 7**). The population may be defined based upon particular characteristics relevant to the exposure or outcome of interest, including geographic, demographic, socioeconomic, or genetic and biological factors. Approaches for identifying the exposure and comparator are discussed by Morgan and colleagues (2018). Research question formulation may also benefit from consideration of the FINER criteria (feasible, interesting, novel, ethical and relevant) (Farrugia et al., 2010).

5. Assembling the Evidence

Because of the diversity of evidence that could be considered in risk decision-making, it is essential to have structured approaches for identifying and summarizing all relevant information. Systematic review provides a powerful approach to meet this need. Guidelines for systematic review of clinical data have been established by the Cochrane Collaboration, including the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009). Although guidelines for best practices in systematic review were developed first for summarizing clinical evidence, similar guidelines are now being developed for other sources of evidence, including toxicology and epidemiology. An overview of best practices in systematic review is provided in a separate article in this issue (Farhat et al., 2020).

Within the field of environmental health in particular, there is a clear need for rigorous systematic review methodologies. Although outside the scope of the present report, less intensive expedited or rapid reviews (see Farhat et al., 2020) may be conducted when time and resources do not permit the completion of a comprehensive systematic review. Review of the scientific evidence plays a critical role in decision-

making about exposures to environmental chemicals by local and federal government agencies. However, challenges exist where there are large data sets, variable study quality, conflicting evidence, or limited information, which impedes integration and final conclusions. Approaches used to assemble and synthesize evidence have evolved over the last three decades, with expert judgement increasingly supported by guidance developed by national and international organizations. Improved methods of chemical assessment that better reflect scientific knowledge have been articulated by the National Research Council in several recent reports (NRC, 2008, 2009, 2014a, 2014b), in particular identifying systematic review as an approach that could substantially improve the processes used to inform policy- and decision-making regarding environmental chemicals.

Several systematic approaches have been evolving and undergoing applications to chemical assessment at the National Toxicology Program (Rooney, Boyles, Wolfe, Bucher, & Thayer, 2014) and U.S. EPA (NRC, 2014a, 2014b). One novel approach to systematic review in environmental health is the Navigation Guide, developed in 2009 through a collaboration between academic scientists and clinicians with the goal of expediting the development of evidence-based recommendations for preventing harmful environmental exposures (Woodruff & Sutton, 2011). The Navigation Guide was developed by drawing from the rigor of systematic review methods used in the clinical sciences with modifications allowing for the unique challenges faced with evidence streams specific to environmental health (i.e., animal toxicology and human epidemiology data).

To date, the Navigation Guide has been applied in five published case studies as proof-of-concept (Johnson et al., 2016; Johnson et al., 2014; Koustas et al., 2014; Lam et al., 2014; Lam et al., 2017; Lam et al., 2016; Vesterinen et al., 2015) (**Table 1**). These case studies were some of the first to demonstrate that systematic and transparent review approaches in environmental health were not only achievable but also advantageous over existing methodologies such as narrative reviews.

The Navigation Guide systematic review methodology involves three main steps:

1) Specify the study question: Frame a specific research question relevant to decision-makers about whether human exposure to a chemical or other environmental exposure is a health risk
2) Select the evidence: Conduct and transparently document a systematic search for published and unpublished evidence
3) Rate the quality and strength of the evidence: Rate the potential risk of bias (i.e., internal validity) of individual studies and the quality/strength of the overall body of evidence based on prespecified and transparent criteria (typically outlined in a pre-published, publicly available protocol). The Navigation Guide methodology conducts this process separately by evidence stream (i.e., human and animal evidence). Ultimately, evidence is combined by integrating the quality ratings of each of these two evidence streams. The end result is one of five possible statements about the overall strength of the evidence: "known to be toxic," "probably toxic," "possibly toxic," "not classifiable," or "probably not toxic."

Table 1.: Five case studies of the Navigation Guide

Citation	Case study	Evidence streams	Findings

Johnson et al. 2014 Koustas et al. 2014 Lam et al. 2014	Developmental exposure to perfluorooctanoic acid (PFOA)and fetal growth outcomes	Human and animal	Rated 18 epidemiology studies and 21 animal toxicology studies. Both evidence streams were rated as "moderate" quality and "sufficient" strength, leading to a final conclusion that PFOA was "known to be toxic" to human reproduction and development.
Vesterinen et al. 2015	Fetal growth and maternal glomerular filtration rates	Human and animal	Rated 31 human and non-human observational studies as "low" quality and two experimental non-human studies as "very low" quality. All three evidence streams were rated as "inadequate." There was insufficient evidence to support the plausibility of a reverse causality hypothesis for associations between environmental exposures during pregnancy and fetal growth.
Johnson et al. 2016	Exposure to triclosan and human development or reproduction	Human and animal	Rated three human studies and eight experimental animal studies in rats reporting hormone concentration outcomes (thyroxine levels). Human studies were rated as "moderate/low" and animal studies were rated as "moderate." There was "sufficient" non-human evidence and "inadequate" human evidence, leading to the conclusion that triclosan was "possibly toxic" to reproductive and developmental health.
Lam et al. 2016	Exposure to air pollution and Autism Spectrum Disorder (ASD)	Human only	Rated 23 epidemiology studies. Evidence was rated as "moderate" quality, leading to the conclusion that there was "limited evidence of toxicity" between exposure to air pollution and ASD diagnosis.
Lam et al. 2017	Developmental exposure to Polybrominated diphenyl ethers (PBDEs) and IQ/ADHD outcomes	Human only	Rated 10 epidemiology studies for intelligence outcomes and 9 studies for ADHD outcomes. Evidence was rated as "moderate" quality with "sufficient" evidence for IQ outcomes and as "moderate" quality with "limited" evidence for ADHD outcomes.

Although various authors have suggested different data hierarchies for use in evidence integration (Burns, Rohrich, Chung, 2012; Petrisor & Bhandari, 2007), consensus on a single hierarchy for application across diverse risk assessment contexts is lacking. As an example, Yetley et al. (2017) identified hierarchies of evidence considering sources of information to support the establishment of dietary reference intakes (DRIs) of nutrients present in the food supply. In this paradigm, well-conducted randomized clinical trials (RCTs) represent the 'gold standard' in terms of obtaining unbiased information directly in human populations. RCTs will not be available for most hazards of concern, though the hierarchy of evidence pyramid shown in **Figure 3** also identifies other valuable sources of information that are frequently used

in risk assessment applications. Other variations of this hierarchy have also been suggested, such as that discussed by Murad *et al.* (2016), which includes consideration of both study design and quality to allow for departures from a strict *a priori* hierarchy of evidence.

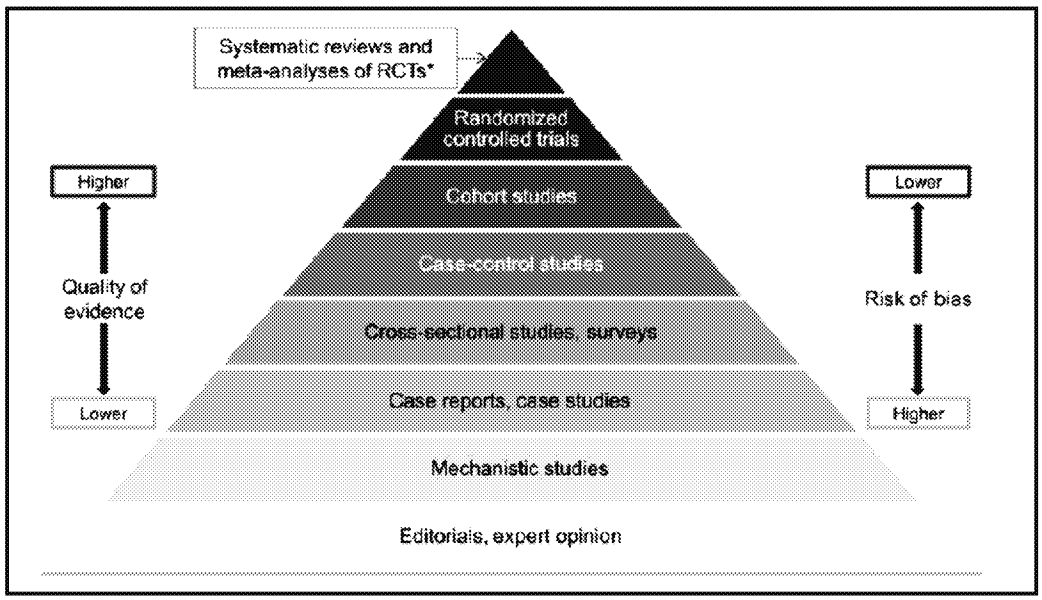


Figure 3. Hierarchy of evidence pyramid
[adapted from Yetley et al. (2017)]

The emphasis on the application of systematic review to assemble the data needed to support evidence-based risk assessment is consistent with recommendations made by the US National Research Council (NRC, 2014a) as part of its review of the US EPA's IRIS program. As indicated in **Figure 1**, adapted from the NRC review, systematic review represents a critical first step in assembling all human, animal, and mechanistic data relevant to the assessment of potential risks associated with environmental health hazards.

Additional guidance on the use of systematic review in risk assessment is available from numerous sources. Farhat and colleagues (2020) trace the evolution of incorporating systematic review into evidence-based risk assessment. Contemporary examples of the application of current methods in systematic review across different domains — including clinical, epidemiological, and toxicological applications — and provide a summary of available tools to support best practices in systematic review.

6. Synthesizing the Evidence

Once all relevant information has been assembled in a systematic review, this information needs to be synthesized, qualitatively and sometimes quantitatively. Qualitative synthesis involves a determination as to whether the exposure of interest constitutes a human health hazard. Within the context of evidence-based risk assessment, this is done using an appropriate evidence integration framework. Should a potential human health hazard be identified through the qualitative synthesis, the next step is to determine whether the available data are sufficient to support an evidence-based quantitative estimate of population health risk.

6.1 Qualitative Synthesis

A number of frameworks for specific types of risk have been developed by different authorities. IARC, for example, has elaborated and refined a well-known scheme for evaluating human carcinogenicity which, over the last 50 years, has led to the identification of 120[5] agents as known causes of human cancer. (To date, only one agent — caprolactam — of the more than 1,000 agents evaluated has been classified as being *probably not carcinogenic to humans*, demonstrating the well-known scientific challenges in establishing a negative outcome with high confidence). It is important to recognize that the IARC framework for evaluating potential cancer risk to humans identifies cancer hazards, but generally does not result in a quantitative estimate of human cancer risk.

Other schema for evaluating non-cancer hazards have also been proposed by other authorities. Rhomberg and colleagues (2013) recently reviewed 50 frameworks in the literature. A preliminary update to this search from 2013 through to the present time is included in a separate publication in this issue (Saunders-Hastings, Rhomberg, & Krewski, 2020). Meanwhile, Martin and colleagues (2018) recently conducted a critical assessment of 24 approaches in an effort to develop a generalized approach.

6.2 Quantitative Synthesis

In many cases, a quantitative estimate of risk will be needed to complete the risk assessment. In the past, quantitative estimates of risk have often been based on identifying a key study (or studies) that are amenable to fitting an appropriate dose-response or exposure-response model. [Although outside the scope of this report, sophisticated analytical techniques, such as Bayesian model averaging (Thomas et al., 2007), can be used to incorporate results from multiple exposure–response models that are compatible with the data.] These models can then be used to develop projections of potential population health risk and associated uncertainty under specified exposure scenarios. These models can also be used to identify a point of departure (PoD) on the dose-response curve that can be used to establish a reference value (RfV) or other toxicity benchmark to serve as a guideline for human exposure (EPA, 2012). The PoD can also be used to establish a margin of exposure (MoE) reflecting the ratio between the toxicity benchmark and estimated or predicted human exposure levels (Thomas et al., 2013).

When multiple studies with quantitative information on dose-response are available, it may be possible to combine the results of these studies. As discussed below, combined analysis of the primary raw data may be possible when the study designs are compatible. When the primary raw data are not accessible for analysis, meta-analysis of summary risk estimates from the individual studies is often done. Another potentially useful approach to combining data involving different toxicological endpoints is categorical regression. Each of these approaches is described briefly below.

6.2.1 Combined Analysis

When access to the primary raw data from a series of related studies is available, a combined analysis of the raw data can be conducted. Having access to the raw data affords maximum flexibility in modelling

[5] https://monographs.iarc.fr/agents-classified-by-the-iarc/

exposure-response relationships across the studies being combined, as well as an opportunity to evaluate the effects of potential modifying factors included in the original studies. For example, Krewski and colleagues (2006) conducted a combined analysis of the primary raw data from a series of case–control studies on residential radon and lung cancer risk, , demonstrating for the first time a strong association between radon and lung cancer in residential settings.

6.2.2 Pooling Epidemiological Data

There are various methods for pooling data from epidemiological studies, each with its own strengths and limitations. Combining primary data from individual studies to yield a large dataset — referred to as *pooled analysis* when used in epidemiological studies — has many advantages. In particular, the increase in sample size allows for more precise calculations of risk estimates (Tobias, Saez, & Kogevinas, 2004). The large sample size can also improve the statistical power to allow the assessment of risks in specific subgroups or restricted subsets of data that would not be possible in smaller data sets of individual studies. These strengths of pooling data make it appealing when investigating effects of rare exposures or risk factors of diseases that have long induction periods, such as cancer (Cardis et al., 2011; Fehringer et al., 2017; Felix et al., 2015; Gaudet et al., 2010; Kheifets et al., 2010; Peres et al., 2018; Wyss et al., 2013).

Although pooling primary sources of data can be expensive and time-consuming and requires the agreement to data-sharing and cooperation of investigators from multiple study centers, the approach does not have the limitations listed above. Pooling primary data can allow investigators to make a broader range of conclusions compared to meta-analyses that are based on published study findings (Checkoway, 1991). This approach also has other advantages; it allows having unified inclusion criteria and definitions of variables across the centers. It also allows the use of the same statistical model on all of the combined data. This is particularly important since individual studies commonly adjust for different confounders in their analysis (Friedenreich, 1993). Standardizing the methods used reduces potential sources of heterogeneity across the studies. Finally, the large sample size allows examination of rare exposures and performing subgroup analyses that would not be feasible in individual studies due to low statistical power.

This type of analysis can be done retrospectively or prospectively. Prospective planning has the added advantage that it allows co-investigators to plan ahead and ensure uniform methods are used for data collection and reporting (Blettner, Sauerbrei, Schlehofer, Scheuchenpflug, & Friedenreich, 1999). Nevertheless, many pooled analyses have been conducted retrospectively after individual studies had reported their findings.

For example, as mentioned above, Krewski and colleagues (2006) retrospectively combined primary data from seven case-control studies in North America. Their pooled findings, based on 4,081 cancer cases, indicate an association between residential radon and lung cancer, although findings from the individual case control studies had provided inconsistent evidence on the risks of lung cancer. The pooling of data further allowed analysis on subsets of the data with more complete radon dosimetry in the most critical exposure time windows. The investigators also performed dose–response analyses based on the histological type of lung cancer (Field et al., 2006). These analyses had not been possible in the prior analyses of each study data individually.

Similar radon risk analyses were conducted by Darby and colleagues (2005) from 13 European case–control studies on 7,148 lung cancer cases. The large sample size achieved by pooling provided sufficient statistical power to detect moderate risks that could not be detected in individual studies.

To perform a pooled analysis of primary data from multiple studies, Friedenreich (1993) suggests the need for a strict protocol. Further, the author details eight steps to follow for pooling data and analysing the combined dataset: 1) identify relevant studies; 2) select (sufficiently similar) studies from which to pool data; 3) combine the data after obtaining each study data from original investigators; 4) estimate study-specific effects using logistic regression; 5) examine the homogeneity of study-specific effects; 6) estimate pooled effects (if study-specific effects are homogenous); 7) explain heterogeneity between studies (if studies-specific effects are not homogenous); and 8) perform sensitivity analyses to examine the robustness of the pooled effects.

6.2.3 Meta-Analysis

Meta-analysis has become a popular and useful technique for developing a more pragmatic estimate of risk by quantitatively combining compatible study-specific risk estimates. Meta-analysis requires that the designs for the studies being combined are reasonably compatible, and that the study results do not demonstrate a high degree of heterogeneity. While meta-analysis has been applied in toxicological risk scenarios, it is predominantly applied in cases where human data is available.

When pooling data, primary data from individual studies are combined to provide a much larger dataset that is then analysed to obtain an overall effect estimate. This is the main difference compared to meta-analyses, where effect estimates reported from individual studies are combined into one overall effect estimate. Meta-analyses are very common in epidemiology, and have many advantages including low associated costs, time efficiency, and the ability to provide an overall quantitative assessment of risk and uncertainty (Friedenreich, 1993). However, limitations do arise in instances where significant heterogeneity between studies is present due to variations in study design, eligibility criteria, exposure and outcome ascertainment and statistical analyses (Blettner et al., 1999). Meta-analyses are also limited by the information provided in the publication and may preclude dose–response analysis and specific subgroup analyses (Friedenreich, 1993; Tobias et al., 2004).

Recent examples of informative meta-analyses of epidemiological data include analyses of the association between exposure to diesel exhaust and lung cancer (Vermeulen et al., 2014) and analyses of the association between talc and ovarian cancer (Taher et al., 2019). Vermeulen and colleagues (2014) conducted a meta-analysis of three epidemiological studies of the association between occupational exposure to diesel exhaust emissions in the mining and trucking industries and lung cancer, using elemental carbon as an indicator of exposure to diesel exhaust. This analysis provides a possible approach to characterizing the exposure-response relationship between diesel exhaust and lung cancer risk (HEI Diesel Epidemiology Panel, 2015). Taher et al. (2019) conducted a meta-analysis of 27 case-control and cohort studies, with limited evidence of study heterogeneity, to estimate the odds for ever use of talc to be 1.28 (95% CI: 1.20–1.37). An important component of this work was the conduct of a series of subgroup

603 analyses, focusing on the nature of talc use, tumor characteristics and the possible effect of menopausal
604 state, hormone use and pelvic surgery.

605 **6.2.4 Categorical Regression**

606 Categorical regression can be used to combine data from diverse sources, including different (both
607 toxicological and epidemiological) types of studies and studies focusing on diverse health endpoints. This
608 is done by developing a severity scoring system to place different adverse health outcomes on a common
609 severity scale, following which categorical regression modelling of the severity scores can be done. In
610 analyses involving both animal and human data, adjustments for inter-species differences and sensitivity
611 can be included in the model.

612 The US EPA has invested considerable effort in developing a software package called CatReg to perform
613 categorical regression (EPA, 2017). More recently, Milton and colleagues (2017a; 2017b) have extended
614 the US EPA CatReg approach to permit modelling of U-shaped dose-response curves for essential
615 elements that demonstrate toxicity due to both excess and deficiency. Yetley and colleagues (2017) have
616 identified categorical regression as a potentially useful tool for combining data from multiple sources in
617 establishing DRIs for nutrients. To illustrate the use of categorical regression in practice, Farrell and
618 colleagues (2020) provide a description of the application of this technique to rich datasets on two
619 essential elements — copper and manganese — that include extensive human and animal data.

620 **6.2.5 Combining Outcomes with Different Severities**

621 Assessment of chemical hazards is based on specific critical health effect(s). As an extension, Sand and
622 colleagues (2018) introduced a method for characterizing the dose-related sequence of the development
623 of multiple (lower- to higher-order) toxicological health effects caused by a chemical. A "reference point
624 profile" was defined as the relation between benchmark doses (BMDs) for selected health effects, and a
625 standardized severity score determined for these effects (**Figure 4**). For a given dose of a chemical or
626 mixture the probability for exceeding the reference point profile can be assessed. Following severity
627 weighing an overall toxicological response (expressed in terms of the most severe outcomes) at the same
628 dose can then be derived by integrating contributions across all health effects. Conversely, dose
629 equivalents corresponding to specified levels of the new response metric can also be estimated. The
630 reference point profile is a cross-section of the dose-severity-response volume, and in its generalized form
631 the method accounts for all three dimensions (dose, severity, response). In this case, the new response
632 metric becomes a proxy for the probability of response for the most severe health effects, rather than the
633 probability for exceeding the BMD for such effects.

634 Conceptually, there are similarities between this method and categorical regression (e.g., Hertzberg and
635 Miller, 1985; Dourson et al., 1997: Milton et al., 2017b). The latter methods have for example been used
636 to calculate the probability for a given severity category. The method introduced by Sand and colleagues
637 (2018) provides this type of output simultaneously across all categories. In addition, as indicated earlier,
638 probabilities are integrated over the entire severity domain to produce an overall response expressed in
639 terms of the most severe health effects. Integration across different severities requires weighting, and a

developed system with nine severity categories (C1 to C9) is therefore mapped to a quantitative severity scale (S = 0 to S = 1) (see **Figure 4**).

Using data from the U.S. National Toxicology Program 2-year studies, Sand and colleagues (2018) demonstrated that results derived by the method are largely insensitive to the choice of model used to describe the reference point profile. The proposed method also appears to be robust with respect to minor and moderate changes in severity classification of BMDs. Further analyses indicate that the interpretation of effective doses or points of departures, based on individual health effects, may change when considering health effects jointly along the lines proposed (Sand, 2020). This influences the consideration of equipotent doses for different chemicals, and the concept of acceptable response levels for individual effects. In addition, results suggest that estimation of exposure guidelines, or similar, by the proposed method may be sufficiently accurate and precise even if data for the most severe health effects, associated with the highest severity categories, are omitted (Sand, 2020). The method may therefore enable derivation of a surrogate for the probability of severe health effects, and/or the probability for exceeding corresponding BMDs, also in the case of using data on comparatively "mild" effects only.

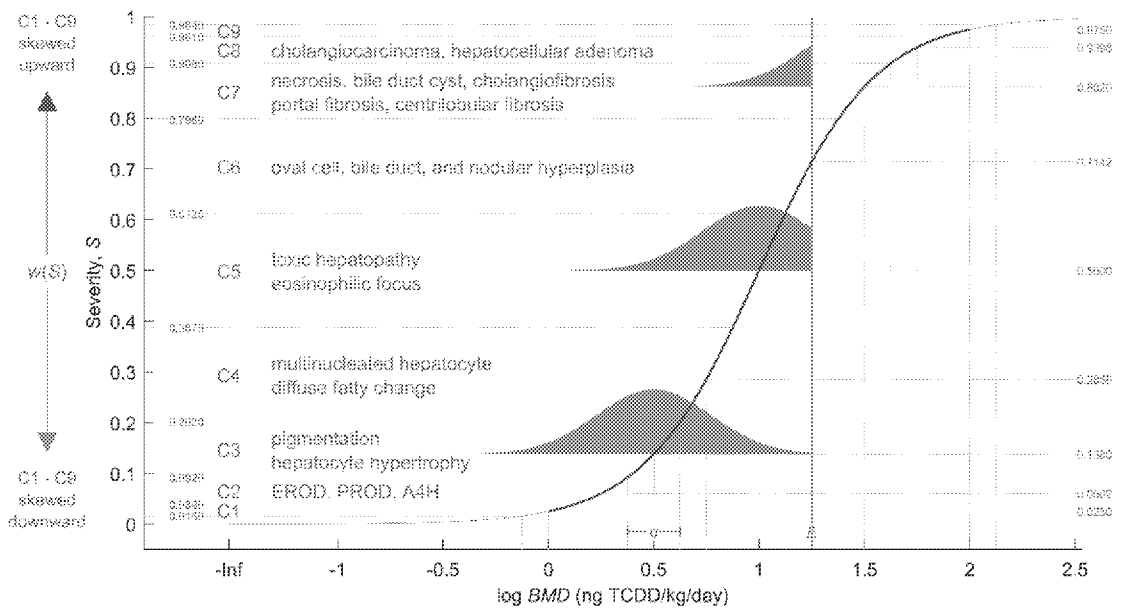


Figure 4. Technical illustration of the reference point profile, which is a cross-section of the dose-severity-response volume. The solid s-shaped (Hill) curve describes the relation between the BMD for selected health effects, and the severity of toxicity (S) determined for these effects. The severity for individual health effects is first determined categorically according to a hierarchical classification scheme: the classification performed by Sand and colleagues (2018) of considered health effect in the liver is illustrated. The nine-graded categorical scale, C1 - C9, is then mapped to a quantitative scale that range from S = 0 to S = 1. The default mapping distributes severity categories symmetrically across S (see Sand et al., 2018 for details). The variability is assumed to be normally distributed on the log-scale with constant variance. Red areas describe probabilities for exceeding the reference point profile at exposure level, E, corresponding to the vertical (red) line. Here, E corresponds to an integrated response of 0.25 (50%), and E intersects the solid curve at $S \approx 0.71$. The midpoint of C6 thus represents the center in terms of the new

response metric. This point of calibration is approximately independent of the model parameters (Sand et al., 2018), and C6 also is regarded as the breaking point between reversible and irreversible effects. Association of C6 to a 50% response is therefore considered as a plausible starting point for severity weighting. A non-linear severity-weight, $w(S) \neq S$, will indirectly modify the default mapping. This allows the midpoint of the system, corresponding to a 50% response, to be lower (C1 - C9 skewed upward) or higher (C1 - C9 skewed downward) than the midpoint of C6, which would also increase or decrease the response associated with E, respectively.

6.2.6 Structured Expert Elicitation

Structured expert elicitation (SEE) is a well-established approach for gauging expert opinion, of particular value in contexts characterized by limited data, low risk and substantial uncertainty (Aspinall, 2008, 2010; Cooke, 2013, 2015). One challenge associated with expert elicitation is the risk or perception of overly subjective expert input. For example, Schünemann and colleagues (2019) point to the need to distinguish expert-elicited *evidence* from expert *opinion*, relying on evidence to inform decision-making. Concerns about expert bias are reduced through an anonymized elicitation procedure with a formal, transparent, and auditable processing of responses and a performance-based weighting scheme for pooling judgements. This encourages experts to be open-minded in responding with their estimates and uncertainties, based on their own personal knowledge, expertise, and experience.

There are nearly 100 specific SEE methodologies, which may concern a generic scenario or circumstances related to the assessment or management for a specific project scenario (Colson & Cooke, 2017; Cooke & Goossens, 2008). Generally speaking, the method quantifies subjective judgements through the weighting of expert responses in order to generate a collective view represented as a median value and accompanying uncertainty distribution. An expert panel between four and twenty members is considered adequate to obtain meaningful results (Colson & Cooke, 2017).

The elicitation may be convened either in person or through video conferencing (thus also reducing the carbon footprint of the event), with experts being offered the opportunity to comment on the process, decline to answer specific questions or withdraw from the exercise entirely. Per Cooke's Classical Model, a SEE should include the following steps (Cooke, 2013, 2015):

- A draft version of the elicitation instrument is reviewed by independent experts (modifications as necessary).
- The elicitation instrument is introduced to the expert participants, with a thorough review of the relevant terms and conditions. Questions are permitted throughout the process, such that the problem, context, definitions, and question content are understood. Refinements may be made, with the goal of ensuring the same understanding among experts.
- Experts complete each question individually using pre-formatted response tables, which are submitted by the experts to the investigators upon completion. Experts first complete a series of "calibration" questions on technical issues for the topic of interest. This calibration exercise enables distinct performance weights to be given to individual experts based on their accuracy and ability to judge uncertainties.

- Experts then respond to numerical uncertainty distribution "target" questions of the same format, with a central value (median), best judgement (50th percentile) and the 90% credible range (lower limit 5th percentile and upper limit 95th percentile).
- Following the analysis, another facilitated meeting or video conference is arranged, providing the expert panel with an opportunity to review preliminary findings. Another round of modification and elicitation may be conducted if necessary.

7. Evidence Assessment & Presentation of Findings

7.1 Determining certainty in the evidence based on the GRADE approach

The evidence identified by systematic reviews can be assessed collectively in a framework such as that described for GRADE (Guyatt, Oxman, Kunz, Atkins, et al., 2011). This moves forward the individual study assessment, so that end-users understand the strengths and limitations (i.e. certainty) across the body of the evidence. Informed by Bradford Hill criteria and the iterative development of evidence-based medicine, the GRADE approach for evidence assessment evaluates the certainty in the evidence based on the following domains that may decrease one's certainty in the body of evidence: risk of bias (i.e. study limitations), inconsistency (i.e. heterogeneity), indirectness, imprecision, and publication bias (Balshem et al., 2011; Guyatt et al., 2011). In addition, for nonrandomized studies, one's certainty of the body of evidence may be increased by the following domains: magnitude of effect (e.g. large or very large effect size), dose-response gradient, or opposing residual confounding (an effect seen in the opposite direction expected from confounders).

These eight domains can help assessors understand the body of the evidence across outcomes as it relates to the research question of interest (i.e. our PECO question, section 3), even when the evidence comes from non-human studies. As mentioned previously, five domains relate to lowering one's certainty in the body of evidence. Risk of bias is informed by the individual study assessments performed as part of the systematic review. While GRADE was originally developed in the context of randomized control trials, its application has expanded to include risk of bias related to randomized and nonrandomized intervention and exposure studies (Guyatt, Oxman, Kunz, Atkins, et al., 2011; Morgan et al., 2019; Schunemann et al., 2019). When considering inconsistency across the pooled evidence, the distinction is made between explained inconsistency or unexplained (Guyatt, Oxman, Kunz, Atkins, et al., 2011). Indirectness, based on how directly the identified evidence answers the research question, is a key element for evidence integration (Guyatt, Oxman, Kunz, Woodcock, et al., 2011). Information used to inform indirectness would be, when the population of interest is humans, how directly can evidence from animal experiments or other types of research (*in vitro* or *in vivo*) be extrapolated or help inform the association between an exposure and outcome. Some research has explored how the domain of indirectness relates to evidence from pre-clinical studies (e.g. research from animals) (Hooijmans et al., 2018). Imprecision considers whether the overall estimate of effect is precise or due to random error. Lastly, publication bias summarizes whether or not all the studies that have been conducted were captured in the review.

When considering the three domains that allow for increased certainty across the body of evidence, the magnitude of effect captures the extent of the observed effect, dose-response considers the exposure-

effect relationship, and opposing residual confounding captures whether or not the worst-case scenario still allows for drawing strong conclusions (Guyatt, Oxman, Sultan, et al., 2011).

Operationalizing these domains relates to the understanding of the relationship between the research question and the evidence extrapolated to inform the findings to that question. In most instances, this can be informed by exploring the various sources of indirectness. For example, one may be interested in humans as a population. For example, humans who are exposed to a carcinogen with the interest in exploring how this exposure would relate to an adverse outcome of interest. Within GRADE, the best available evidence is understood to come from human studies; however, indirect evidence from other sources (animal, mechanistic) is also considered. The exposure or comparator could also introduce an element of indirectness. To apply this to a review, five paradigmatic scenarios exist (Morgan et al., 2018). In one particular scenario, little is known about the association between the exposure and outcome, therefore the assessment seeks to define that relationship. In this situation, mechanistic data or modelling may be utilized to see whether or not we have some confidence in a statement between the exposure and outcome. Indirectness may also be identified within the outcome, as to whether or not out evidence is extrapolated from a surrogate.

The starting point when determining the certainty across the body of evidence for an outcome typically starts at high certainty for RCTs and low certainty for nonrandomized studies. However, with the development of risk of bias instruments applied to nonrandomized studies that use a standardized scale from RCTs, the level of certainty could be increased (Morgan et al., 2019; Schunemann et al., 2019).

7.2 Relevance to Risk Assessment

Characterizing the relationship between exposure levels and the health impacts they exert has been the focus of much research conducted by risk assessors, health practitioners, and regulatory experts in developing health protection programs to establish safe intake levels for humans (Krewski et al., 2010; Stern et al., 2007). Any substance, including but not limited to chemicals, nutrients, vitamins, or pharmaceuticals, has the potential to be harmful to humans if they are exposed to too much or too little. To establish a range of allowable intake for a substance that may be harmful to humans in excess or deficient amounts, it is necessary to strike a balance between the health impacts exerted by exposures across the excess-deficiency spectrum. The challenge of identifying an acceptable medium between excess and deficiency motivates the development of exposure–response models; they provide the foundation for identifying recommended levels of exposure to essential and nonessential substances (Krewski et al., 2010).

The US EPA developed the concept of a RfV for toxic substances, which has been widely accepted and used in practice to prescribe safe intake levels for humans. The RfV can be derived from a single key study that considers one critical health effect and is defined by applying uncertainty or adjustment factors to the no-observed-adverse-effects level (NOAEL), which corresponds to the level of exposure that does not result in a significant increase in the risk of adverse effects in the exposed group when compared with controls (Barnes & Dourson, 1988). Currently, the benchmark dose (BMD) is more often used as the basis for determining the RfV. The BMD was introduced by Crump (1984) and uses a mathematical model to

identify the dose corresponding to a specified increase in response. It has also been used as a point of departure on the dose–response curve for establishing human exposure guidelines (European Chemicals Agency, 2013). More recently, the signal-to-noise crossover dose (SNCD), defined as the dose at which the uncertainty in the biological signal is indistinguishable from the background noise, has been introduced as an alternative to the BMD (Sand, Portier, & Krewski, 2011). Categorical regression has also been applied in dose–response modelling for health risk assessment (Allen, Zeiger, Lawrence, Friedman, & Shipp, 2005; Chambers et al., 2010; Gift, McGaughy, Singh, & Sonawane, 2008; Haber, Strickland, & Guth, 2001; Milton et al., 2017a; Milton et al., 2017b). Categorical regression facilitates the inclusion of multiple studies in an exposure–response model by applying a severity scoring scheme to standardize different outcomes reported in each experiment. Further discussion of these techniques may be found in previous publications (Milton et al., 2017a; National Research Council (US) Committee on Improving Risk Analysis Approaches Used by the US EPA, 2009; National Research Council (US) Committee on Risk Assessment of Hazardous Air Pollutants, 1994; Yetley et al., 2017).

8. Uncertainty Analysis

8.1 Assessing Uncertainty in Risk Assessment

In a great majority of situations of interest, risk assessment is confronted with a wide range of imperfections in the evidence base that collectively constrain the ability to provide a certain answer with respect to causality, and even if causality is assumed, a certain estimate of the level of risk to be expected in the exposed population of interest. The level of uncertainty in the level of risk can span multiple orders of magnitude, even when excluding uncertainty in causality.

In general, guidance documents that prescribe best practices in risk assessment suggest that a formal treatment of uncertainty is a fundamental component of risk assessment processes (NRC, 2009). The sources of uncertainty (i.e., the limitations of the evidence based) can be from a wide spectrum of sources, including:

- inability to infer a causal relationship due to conflicting evidence within the same line of evidence;
- inability to infer a causal relationship due to conflicting evidence from different lines of evidence;
- lack of testing data for numerous types of possible health outcomes;
- testing data which is available but of questionable relevance to human health;
- uncertainty in what to predict at low doses given data restricted to much higher doses;
- uncertainty in the fate and transport of substances and the resulting environmental concentrations in near- and far-field exposure situations;
- the levels of exposure in the population of interest to the environmental media or product in question;
- uncertainty stemming from imprecision of knowledge of mechanisms of action; and
- uncertainty stemming from imprecisions in goals such as the percentile of the population to be protected by an exposure threshold value.

816 Each of these forms of uncertainty can be dealt with in a piece-wise fashion with the uncertainty
817 characterization limited to an interim conclusion within different stages of the risk assessment process. In
818 addition, the means of characterizing uncertainties can vary from a narrative approach (explaining the
819 limitations or extent of doubt of any conclusions and the basis for that doubt), to qualitative (labelling
820 uncertainty as high, medium, low) to categorical (e.g., causal, likely to be causal, suggestive evidence of
821 causal, etc.) through to quantitative characterizations (providing a probability estimate for causality,
822 providing a confidence interval on an estimated value, or explicitly defining or computing a probability
823 distribution representing uncertainty in an estimated or computed value).

824 The various practices employed determine, partly, to what extent a final and overall characterization of
825 uncertainty can be rendered, even if the uncertainties in each part of the risk assessment are relatively
826 complete. Fully quantitative estimates of uncertainty in each part of a risk assessment can be propagated
827 in a reliable way to capture the overall uncertainty. However, it is currently rare that practitioners capture
828 uncertainty in a quantitative way in all the ways that the inadequacies of the evidence base might
829 contribute to the overall uncertainty in a risk assessment.

830 For the three main risk assessment steps preceding risk characterization, recent methods have emerged,
831 some have long existed and other methods need to be developed to allow for the careful and
832 comprehensive treatment of uncertainty in risk assessment.

833 In hazard identification, progress is being made in minimizing some sources of uncertainty due to the
834 increasing use of formal systematic methods in the gathering and treatment of evidence (e.g., explicit
835 criteria for inclusion and quality scoring of studies) within the principal lines of evidence. However,
836 methods to formally capture the uncertainty and implications of imperfections in the evidence base when
837 integrating evidence across evidence lines appears to have limited formal methodological support, often
838 relying on expert judgement and consensus-based processes in the ultimate weighing of evidence.

839 One potential method of formally weighing and combining evidence across evidence lines was described
840 during the workshop in a proof-of-concept application. This method captures the imperfections of each
841 type of evidence (both in a general sense as well as in a study-by-study sense) in the form of a Conditional
842 Probability Tables linking *evidence* of various types and qualities to *hypotheses* in a Bayesian Network. The
843 Bayesian Network is a means of computationally combining all the evidence, through application of Bayes'
844 Rule within the software tool, to yield intermediate and overall statements of the uncertainty in a
845 hypothesis linked to the various sources of evidence in the Network. The explicit linkages captured in the
846 Network allow for detailed sensitivity analysis both in a general sense (how strongly is an animal-based
847 test of type X linked to evidence of health outcomes Y) and in a specific sense (how strongly does the
848 overall conclusion of carcinogenicity for substance A depend on the quality score applied to study Z?).

849 For exposure assessment, with some effort, quantitative estimates of uncertainty can be readily combined
850 (e.g., through Monte Carlo simulation or other methods of propagating uncertainty) from the
851 uncertainties in individual quantitative components (contaminant levels, food intakes, product use,
852 occupational working conditions, inhalation rates, body weights) that are combined to generate estimates

of dose. While non-trivial, conducting an uncertainty assessment within the step of exposure assessment is supported by existing methods.

Dose-response assessment is an area where recent advances may be considered the most dramatic. The availability of software tools to enable dose-response assessment, including capturing uncertainty in the process, has been an important contribution. Efforts to harmonize non-cancer and cancer dose-response assessment into a formal and quantitative process have made significant progress and are now widely applicable with supporting publicly available tools (see 8.2 below; Chiu et al., 2018). The ability to consider varying levels of severity in a quantitative way is another important consideration in removing uncertainty that stems from the relative vagueness in the treatment of severity has been considered in the traditional minimally quantitative dose-response approaches (see 6.2.5 above; Sand et al., 2018).

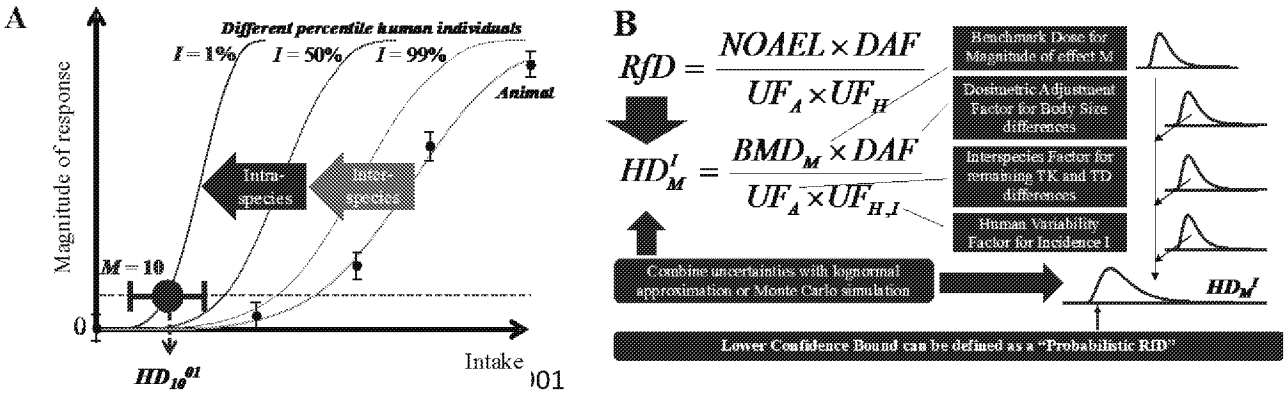
The risk characterization step of risk assessment integrates insights from the hazard identification, exposure assessment and dose-response steps to generate overall estimates of risk and uncertainty surrounding those estimates. With recent advances in the treatment of dose-response assessment, the ability to render a more complete characterization of uncertainty in risk estimates should be enabled. The inclusion of uncertainty in causality aspects of hazard identification in an integrated way may be the most challenging to render and to communicate.

8.2 The IPCS approach to quantitative uncertainty analysis

As noted above, the RfV and similar approaches to quantitative synthesis suffer from some key limitations, many of which have been identified by several National Academies reports (National Research Council (US) Committee on Improving Risk Analysis Approaches Used by the US EPA, 2009; National Research Council (US) Committee on Risk Assessment of Hazardous Air Pollutants, 1994). For instance, the way in which RfV have traditionally been derived characterize neither the degree of residual risk that may be present nor the shape of the dose-response curve for adverse effects. This is because the RfV is usually obtained by identifying a NOAEL of intake from an experimental study (predominantly animal studies), and then dividing this intake level by a number of "uncertainty factors" to account for limitations in the data. The most commonly applied uncertainty factors are a factor of 10 to address differences between experimental animals (UF_A) and a second factor of 10 to address variability among humans (UF_H). This "NOAEL divided by 100" concept dates back to 1950s in the context of FDA regulation of food additives (Lehman, 1954). Each of these components — the NOAEL, UF_A, and UF_H — is assumed to be "conservative" in the sense of erring on the side of protecting public health, but without much specificity as to "how conservative" they actually are (WHO/IPCS, 2014). For instance, with respect to the NOAEL, it is assumed that the severity of effects at this exposure level are negligible, but the extent to which this is true depends on the endpoint examined and the statistical power of the study (Crump, 1984; EPA, 2012). For UF_A, it is assumed that humans are generally no more than 10-fold more sensitive than the experimental animal species, but it is unclear at what confidence level this 10-fold factor is supposed to be (90%, 95%, 99%?). Similarly, for UF_H, it is assumed that individuals more susceptible to toxicity are no more than 10-fold more sensitive than more typical individuals. Here, there are two ambiguities: first, like UF_A, the confidence level of this 10-fold factor is unclear; 90%, 95%, 99%? Second, it is unclear what

"susceptible" means in terms of the more sensitive tail of the population distribution; 5%, 1%, 1 in a million?

More recently, the World Health Organization/International Program on Chemical Safety (WHO/IPCS) had developed a guidance document describing a "probabilistic" framework that results in substantially better characterization of the intake-response for adverse effects (W. A. Chiu & Slob, 2015). The key concept underlying the WHO/IPCS approach is that the goal of deriving quantities like the reference dose (RfD) is a "target human dose" HD_M^I, defined as to estimated human dose (or intake) at which effects with magnitude M occur in the population with an incidence I, along with an associated confidence interval (**Figure 5A**).



"Target human dose," HD_M^I, defined as to estimated human dose (or intake) at which effects with magnitude M (e.g., 10) occur in the population with an incidence I (e.g., 1%), along with an associated confidence interval.

Figure 5. Summary of probabilistic approach to characterize quantitative uncertainty and variability in dose-response assessment. A) Illustration of the concept of the "human target dose," HD_M^I, which replaces traditional toxicity values such as the Reference Dose (RfD). B) General approach to deriving HD_M^I probabilistically, in comparison to deriving a traditional deterministic RfD. For additional details, see WHO (2014) and Chiu and Slob (2015).

The derivation of the HD_M^I can be summarized into the following steps (illustrated in **Figure 5B**):

- **Replace the NOAEL with so-called "benchmark dose (BMD) modeling."** The limitations of the NOAEL as a starting point for toxicological risk assessment have been recognized for decades (European Food Safety Authority, 2015; National Academies of Science, 2001; EPA, 1995, 2012; WHO/IPCS, 2009), and it is generally accepted that the BMD, introduced by Crump (1984), is more scientifically appropriate. The BMD is the dose associated with a specific size of effect, the benchmark response (BMR) [we use M, for magnitude of effect, to denote this value]. The BMD_M is estimated, with associated confidence interval/statistical distribution, by statistical model fitting to dose-response data. In this way, it provides information that the NOAEL does not regarding the nature and severity of the adverse effects under consideration, as well as the precision with which the associated intake level can be estimated.

- **Replace interspecies UF_A with results from mathematical/computational modeling.** The approaches for replacing UF_A all involve disaggregating it into two components. The first component is related to dosimetry, denoted as dosimetric adjustment factor (DAF), to convert experimental animal exposures to "human equivalent" exposures. If chemical-specific toxicokinetic (TK) data are available, PBPK models have been used to derive the DAF (Corley et al., 2012; Dorman et al., 2008; Schroeter et al., 2008; Teeguarden, Bogdanffy, Covington, Tan, & Jarabek, 2008). Otherwise, generic approaches have been developed based on physiological differences, such as allometric scaling by body mass, between experimental animals and humans (EPA, 1994, 2011; West, 1999). A second component, which is denoted UF_A, accounts for any remaining (i.e., unknown) chemical-specific interspecies differences, such as due to toxicodynamics (TD). WHO/IPSC reviewed analyses of historical data on toxicity thresholds across species and chemicals, deriving confidence intervals for the DAF and remaining interspecies differences for oral intakes (WHO/IPCS, 2014).
- **Replace human variability UF_H with results from mathematical/computational modeling.** In the WHO/IPCS framework, UF_H is replaced with a value that depends on the I^{th} population percentile of susceptibility, denoted $UF_{H,I}$ (Chiu & Slob, 2015; WHO/IPCS, 2014). This quantity reflects TK and TD differences between individuals at the median and the I^{th} percentile of the population distribution. For TK, PBPK models have been used for specific substances to estimate the degree of human variability. WHO reviewed previous analyses of historical data on TK and/or TD variability across chemicals (WHO/IPCS, 2014), based on work by Hattis and colleagues (Hattis, Baird, & Goble, 2002), and have recommend "default" factors (as probability distributions) that can be applied.
- **Combine the newly defined BMD, UF_A, and $UF_{H,I}$ in a probabilistic manner to derive an intake-response function and its uncertainty.** The integration of BMD modeling, allometric scaling, and historical TK/TD variability data across chemicals leads to the HD_M^I (Chiu & Slob, 2015; WHO/IPCS, 2014). The HD_M^I can then replace the UL and disaggregates "risk" into the distinct concepts of magnitude of effect (M), incidence of effect (I), and uncertainty (reflected in the confidence interval). The HD_M^I can furthermore be mathematically "inverted" to derive an intake-response function for a specified fraction I of the population (Chiu & Slob, 2015).

The resulting output is a two-dimensional distribution of intake-response functions: one reflecting human variability in terms of I^{th} percentiles, and the other reflecting statistical uncertainty. By providing intake-response functions rather than "bright lines," changes in risk of adverse effects from changes in dose can be quantified. This type of "risk-benefit" comparison would be infeasible under the traditional "NOAEL divided by 100" approach, because there is no characterization of the gradient of the dose-response over a wide enough range of doses. However, the approach to derive an HD_M^I would enable such comparison to be made much more easily. Chiu and colleagues (2018) recently applied this approach to over 600 chemicals and 1,500 endpoints, demonstrating the feasibility of broadly implementing this approach in chemical risk assessments.

9. Preliminary Evidence-Based Risk Assessment Framework

The overarching goal of this project was to develop an evidence-based risk assessment framework to guide the conduct of evidence-based risk assessment, including assembling and synthesizing all relevant data as discussed above. This document provides detailed guidance on each of the key steps involved in conducting an evidence-based risk assessment and is of value to practitioners seeking to ensure that risk

963 assessments are completed according to the highest possible scientific standards, and are conducted in
964 an open, transparent, and reproducible manner.

965 An important aspect of the framework is the distinction between the related steps of assembling and
966 synthesizing the evidence. Systematic review offers a powerful approach to assembling all relevant data
967 in support of the assessment, with objective inclusion/exclusion criteria and study quality assessment. In
968 the past, the selection of studies to be included in risk assessments has sometimes been a source of
969 controversy. By invoking current best practices in systematic review, it is expected that much of this
970 controversy can be circumvented.

971 Having agreed on the evidence base to support risk assessment, attention can then focus on qualitative
972 and possibly quantitative syntheses of the available information. At this stage, clear criteria for evaluating
973 the available data will serve to support data-driven determinations regarding the existence or otherwise
974 of a human health hazard. Should a human health hazard be identified using the criteria embodied in the
975 framework, methods for quantitative syntheses of the available data can then be applied, in cases where
976 the available data are sufficient to support an evidence-based estimate of potential population human
977 health risk and associated uncertainty.

978 An initial framework for the evaluation of all available evidence on the association between a particular
979 exposure and adverse outcome is presented in **Figure 6**. Although practitioners will be familiar with each
980 of the components of this framework, it may serve as a useful paradigm for ensuring consistency in
981 evidence-based risk assessment. Following problem formulation (an important starting point, but outside
982 the scope of the present framework), the first step is to assemble all relevant evidence relating to the
983 specific risk issue under consideration. As discussed, systematic review provides a powerful set of
984 methodologies for accomplishing this in a comprehensive and reproducible manner.

985 In the framework laid out in **Figure 6**, once all relevant evidence has been summarized, the volume of
986 evidence is assigned to one of three tiers: data-poor (Tier 1), limited data (Tier 2), or data-rich (Tier 3). In
987 the data-poor context, a decision is required as to whether the assessment should proceed to a formal
988 evaluation: should the data be judged inadequate to support a meaningful evaluation, key data gaps
989 should be identified and filled before proceeding.

990 Within the context of the present framework, the term 'limited data' (Tier 2) is used to represent at least
991 a minimal amount of data that would support a credible risk assessment. The term 'data-rich' (Tier 3)
992 represents the case in which considerable data is available from multiple sources (including human,
993 animal, and other experimental sources) to support a credible evaluation.

994 Once it has been determined that the data are adequate to support assessment development (regardless
995 of whether the available evidence falls into Tier 1, 2 or 3), the next step is to conduct a qualitative synthesis
996 of the available data, resulting in a determination as to whether or not a human health hazard exists. This
997 is a non-trivial undertaking and will involve the application of explicit (likely context-specific) criteria for
998 hazard determination.

Although such criteria have been elaborated for certain outcomes such as cancer (specifically those elaborated upon by IARC in the January 2019 update to the IARC monographs) (Samet et al., 2019), other criteria will need to be developed for other adverse outcomes. Lessons may be learned from current REACH guidance, which focuses on 11 broad adverse outcomes (Armstrong et al., 2020). The end result of the qualitative synthesis is a statement about the evidence for a causal association between the exposure and outcome(s) of interest. Should an inconclusive outcome be reached, outstanding data gaps should be noted and addressed for use in a future re-evaluation. In its review of the U.S. EPA IRIS program, the NRC recommended categories of evidence that included "sufficient to infer a causal relationship", "suggestive but not sufficient to infer a causal relationship", "inadequate to infer the presence of a causal relationship", and "suggestive of no causal relationship (NRC, 2014, p.94). Other investigators have also suggested simplified evidence categorization schemes, including Wigle and colleagues (2008) and Krewski and colleagues (2017) who proposed categories for classifying evidence as "sufficient", "limited" or "inadequate".

Should the qualitative synthesis conclude that a human health hazard exists, a quantitative synthesis of the available data can be attempted, with the goal of characterizing the level of risk, and attendant uncertainty, in quantitative terms. Recent trends in data aggregation have provided powerful new approaches to quantitative data synthesis, including techniques such as categorical regression that permit the inclusion of quantitative data from multiple sources and on multiple endpoints into a single dose-response analysis. It is important to note that the successful completion of a qualitative synthesis of the available data does not guarantee that the data can support a meaningful quantitative synthesis; a quantitative synthesis will only be possible when there is reliable data on the dose–response relationship between the agent and outcome of interest from one or more sources.

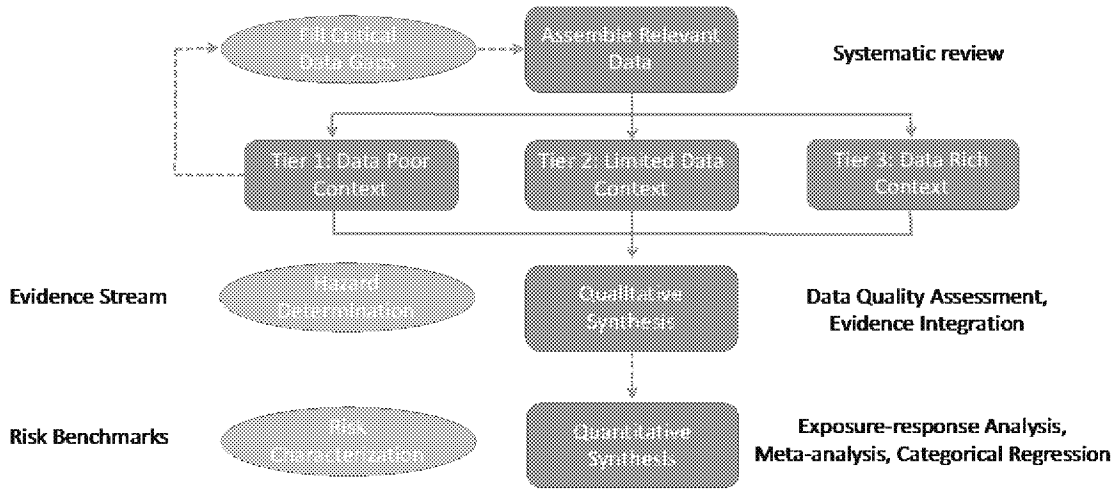


Figure 6. Preliminary evidence-based risk assessment framework

10. Summary of Workshop Deliberations

Experts from the field of evidence integration met at a workshop hosted by the University of Ottawa on December 17–18, 2018 to discuss the development of an evidence-based risk assessment framework. The workshop benefitted from strong attendance from government, academia and industry.

The workshop agenda (provided in full in the **Supplementary Material**) covered a range of topics relevant to risk science and evidence integration, from emerging trends and new methodologies to approaches for putting evidence integration into practice.

During the second day of the workshop, participants were divided into breakout groups, and tasked with addressing four key questions relating to the development of a framework for evidence-based risk assessment. The four discussion areas, described in additional detail below, were:

1. Lessons learned from previous experience
2. Benchmarks of good practice
3. Problem formulation and data requirements
4. Potential challenges

10.1 Breakout Group 1: Lessons Learned from Previous Experience

This breakout group explored what can be learned from previous experience in the development of an organizing framework for evidence integration. Participants explored past successes and challenges, as well as opportunities for future improvement.

The iterative development and refinement of systematic and transparent approaches for data collection, screening and abstraction were identified as key areas of strength where insights could be drawn from past experience. Participants noted, however, that guidance tends to be less clear for mechanistic data. Also, there was a discussion of the need to preserve evidence and decision context throughout the risk assessment process.

It was also suggested that the ability to both summarize individual lines of evidence and use insights from one line to inform judgements for another had improved over time. This was driven in part by best practices in the clinical and pharmaceutical industry and participants suggested that similar efforts for evidence-based risk assessment could benefit from building upon this foundation.

Participants noted that "getting the data right does not mean getting the answer right". In other words, while a systematic approach is essential, both subject matter expertise and public review should be included as necessary components of the risk assessment process. Similarly, transparency should not be conflated with objectivity, and both decisions and their impacts on the assessment process should be made explicit.

Moving forward, participants recommended a focus on developing discrete steps for evidence integration that record and summarily present choices, assumptions and justifications. They also recommended the development of reporting guidelines for author publication of individual studies to facilitate their

incorporation in risk assessments. Lastly, participants advocated for continued knowledge-sharing and collaboration in order to foster best practices and reduce duplication of work.

10.2 Breakout Group 2: Benchmarks of Good Practice

This breakout group explored examples of previous evidence integration efforts in an effort to identify examples of best practices. Participants sought to address this topic by identifying both useful frameworks and relevant case studies to examine their application in practice.

With regard to applicable frameworks, the participants made note of several of the frameworks discussed in earlier sections of this workshop report. One framework that received particular attention was the updated Whaley literature review appraisal toolkit, which participants suggested could be useful as a mechanism for refining best practices and evaluation criteria for systematic reviews.

The case studies discussed were intended to explore various systematic review frameworks, highlighting strengths and areas that could be improved moving forward. Participants compared numerous case studies, such as OHAT immunotoxicity (involving human, animal and mechanistic data), Health Canada data-poor case studies, and EPA IRIS chemical assessments.

From this discussion, participants developed three key recommendations for future consideration.

1. Prospective case studies could be considered as an avenue for advancing discussion of best practices
2. Empirical studies of how to reduce risk of bias could further increase the confidence with which risk assessment findings could be interpreted
3. Efforts to evaluate the impact of different levels of literature review comprehensiveness could be informative in guiding best practices and pragmatism in review efforts

10.3 Breakout Group 3: Problem Formulation and Data Requirements

This breakout group explored how problem formulation can be used to define data requirements for evidence-based risk assessment, as well as how contextual factors may impact these requirements. Participants strongly agreed that risk assessment data requirements are context-dependent and focused their discussion on the use of GRADE to respond to research questions with varying degrees of urgency.

Participants examined how GRADE can be applied to assess research questions across different timescales, including emergency response (within hours), urgent response (1–2 weeks), rapid response (1–3 months) and routine response (more than three months). The types of evidence incorporated, and risk of bias associated with findings would vary across these categories, demonstrating the context-specific nature of data requirements (and flexibility of the GRADE process). While GRADE principles can be applied to results of both systematic and expedited reviews, non-systematic reviews may often lead to greater uncertainty in the interpretation of review findings.

10.4 Breakout Group 4: Potential Challenges

This breakout group discussed what challenges may arise in efforts to strengthen evidence-based risk assessment. Participants considered risk assessment within the broad categories of evidence retrieval (data collection), integration within a stream of evidence and integration across evidence streams.

Within the context of evidence retrieval, participants highlighted the need to be comprehensive without diluting the findings of the review. They identified the need to improve the quality of evidence that informs risk assessment, and identified knowledge translation, data sharing and methodological standardization as possible avenues for overcoming this challenge.

In assessing individual lines of evidence, participants discussed challenges relating to the weighting of difference sources (such as among several animal models) and reaching agreement on the "quality" of information available. Specifically, they discussed what limitations should be viewed as more problematic than others.

When integrating evidence across evidence streams, participants discussed challenges driven by uncertainty about the informative value of different evidence. Integration of qualitative and quantitative data can be particularly challenging in this regard, and participants proposed standardized metrics and probabilistic approaches as potential avenues for overcoming these difficulties.

At a general level, authors suggested that key challenges included building confidence in the risk assessment process, synthesizing different types of evidence, and navigating inter-reviewer disagreements and judgements. They suggested that complementary evidence could be used to support generalizations of risk assessment findings, and that evidence integration at different levels (e.g., to understand the biological plausibility of mechanistic data) could be of value.

11. Conclusion

The overarching objective of this initial workshop was to explore the development of an evidence-based framework for risk assessment. Participants at the workshop discussed issues relating to recent advances in risk science, new methodologies in evidence evaluation, approaches for qualitative and quantitative evidence synthesis, and putting evidence integration into practice.

A preliminary framework was distributed in advance of the workshop and refined based discussion and debate among the workshop participants. The framework presents a practical approach to evidence evaluation and synthesis designed to ensure that the relevant evidence for human health risk assessment is considered in a comprehensive and objective manner. The framework relies on current best practices in systematic review to summarize human, animal and experimental evidence relevant to the risk issue under consideration. With the recent advances in systematic review methodology and powerful software to support the conduct of systematic review, the relevant evidence can be readily summarized in a comprehensive and reproducible manner.

Once the available evidence has been summarized, evidence can be used to conduct a qualitative synthesis of the available data. Application of appropriate integration approaches can be used for hazard

determination. If a human health hazard is identified, a quantitative synthesis of the data can then be undertaken to characterize population health risks and uncertainties in quantitative terms. Both the qualitative and quantitative syntheses will require adequate data to support these syntheses, with key data gaps being identified and filled as relevant to the context of the risk assessment.

Elaborating on the evidence-based risk assessment framework proposed here will require more in-depth consideration of the criteria to be applied in conducting both qualitative and quantitative synthesis of the data. Such criteria will be proposed in a subsequent phase of this work. Upon completion of the evidence-based risk assessment framework, case study prototypes could be conducted to evaluate its use in practice. A follow-up workshop is currently being planned to flesh out the preliminary framework for evidence-based risk assessment in more detail.

Acknowledgements

Any opinions expressed in this article are those of the author(s) and do not, necessarily, reflect the views, official positions, or policies of the authors' institutions. The authors are grateful to Dr. Sue Gapstur, Dr. Marc Beal and Dr. Benny Ling for their helpful comments on the penultimate draft of this manuscript. D. Krewski is the Natural Sciences and Engineering Research Council of Canada Chair in Risk Science at the University of Ottawa.

11. References

Aiassa, E., Merten, C., & Martino, L. (2020). EFSA's Framework for Evidence-Based Scientific Assessments: A Case Study on Uncertainty Analysis. *ALTEX, under review*.

Allen, B., Zeiger, E., Lawrence, G., Friedman, M., & Shipp, A. (2005). Dose-response modeling of in vivo genotoxicity data for use in risk assessment: some approaches illustrated by an analysis of acrylamide. *Regul Toxicol Pharmacol, 41*(1), 6-27. doi:10.1016/j.yrtph.2004.09.006

Anastas, P. T., Sonich-Mullin, C., & Fried, B. (2010). Designing science in a crisis: the Deepwater Horizon oil spill. *Environ Sci Technol, 44*(24), 9250-9251. doi:10.1021/es103700x

Andersen, M. E., McMullen, P. D., Phillips, M. B., Yoon, M., Pendse, S. N., Clewell, H. J., . . . Clewell, R. A. (2019). Developing context appropriate toxicity testing approaches using new alternative methods (NAMs). *ALTEX, 36*(4), 523-534. doi:10.14573/altex.1906261

Armstrong, V., Karyakina, N., Nordheim, E., Arnold, I., & Krewski, D. (2020). Intent and provision of REACH: an overview. *Neurotoxicology, submitted.*

Aspinall, W. (2008). Expert judgment elicitation using the Classical Model and EXCALIBUR. In *Briefing notes for seventh session of the Statistics and Risk Assessment section's international expert advisory group on risk modeling: Iterative risk assessment processes for policy development under conditions of uncertainty/emerging infectious diseases: Round IV*. Retrieved from http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf

Aspinall, W. (2010). A route to more tractable expert advice. *Nature, 463*, 294-295. doi: 10.1038/463294a.

Balshem, H., Helfand, M., Schunemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., . . . Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol, 64*(4), 401-406. doi:10.1016/j.jclinepi.2010.07.015

Barnes, D. G., & Dourson, M. (1988). Reference dose (RfD): description and use in health risk assessments. *Regul Toxicol Pharmacol, 8*(4), 471-486. doi:10.1016/0273-2300(88)90047-5

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T., & Friedenreich, C. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol, 28*(1), 1-9. doi:10.1093/ije/28.1.1

Burns, P. B., Rohrich, R. J., & Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery, 128*(1), 305-310. doi:10.1097/PRS.0b013e318219c171

Cardis, E., Armstrong, B. K., Bowman, J. D., Giles, G. G., Hours, M., Krewski, D., . . . Vrijheid, M. (2011). Risk of brain tumours in relation to estimated RF dose from mobile phones: results from five Interphone countries. *Occup Environ Med, 68*(9), 631-640. doi:10.1136/oemed-2011-100155

Chambers, A., Krewski, D., Birkett, N., Plunkett, L., Hertzberg, R., Danzeisen, R., . . . Slob, W. (2010). An exposure-response curve for copper excess and deficiency. *J Toxicol Environ Health B Crit Rev, 13*(7-8), 546-578. doi:10.1080/10937404.2010.538657

Checkoway, H. (1991). Data pooling in occupational studies. *J Occup Med, 33*(12), 1257-1260. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1800686

Chiu, W., Axelrad, D., Dalaijamts, C., Dockins, C., Shao, K., Shapiro, A., & Paoli, G. (2018). Beyond the RfD: Broad Application of a Probabilistic Approach to Improve Chemical Dose-Response Assessments for Noncancer Effects. *Environmental Health Perspectives, 126*(6), 067009-067009. doi:10.1289/EHP3368

Chiu, W. A., & Slob, W. (2015). A Unified Probabilistic Framework for Dose-Response Assessment of Human Health Effects. *Environ Health Perspect, 123*(12), 1241-1254. doi:10.1289/ehp.1409385

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety, 163*, 109-120. doi:https://doi.org/10.1016/j.ress.2017.02.003

Cooke, R. M. (2013). *Validating Expert Judgment with the Classical Model*. Retrieved from http://www.expertsinuncertainty.net/LinkClick.aspx?fileticket=HlcTmEoDunY%3D&tabid=4385&mid=8296

Cooke, R. M. (2015). The aggregation of expert judgment: do good things come to those who weight? *Risk Anal, 35*(1), 12-15. doi:10.1111/risa.12353

Cooke, R. M., & Goossens, L. L. H. J. (2008). TU Delft expert judgment database. *Reliability Engineering and Systems Safety, 93*, 657-674. doi:10.1016/j.ress.2007.03.005

Corley, R. A., Kabilan, S., Kuprat, A. P., Carson, J. P., Minard, K. R., Jacob, R. E., . . . Einstein, D. R. (2012). Comparative computational modeling of airflows and vapor dosimetry in the respiratory tracts of rat, monkey, and human. *Toxicol Sci, 128*(2), 500-516. doi:10.1093/toxsci/kfs168

Cote, I., Andersen, M. E., Ankley, G. T., Barone, S., Birnbaum, L. S., Boekelheide, K., . . . DeWoskin, R. S. (2016). The Next Generation of Risk Assessment Multi-Year Study-Highlights of Findings, Applications to Risk Assessment, and Future Directions. *Environ Health Perspect, 124*(11), 1671-1682. doi:10.1289/ehp233

Crump, K. S. (1984). A new method for determining allowable daily intakes. *Fundam Appl Toxicol, 4*(5), 854-871. doi: 10.1016/0272-0590(84)90107-6.

Darby, S., Hill, D., Auvinen, A., Barros-Dios, J. M., Baysson, H., Bochicchio, F., . . . Doll, R. (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *Bmj, 330*(7485), 223. doi:10.1136/bmj.38308.477650.63

Dorman, D. C., Struve, M. F., Wong, B. A., Gross, E. A., Parkinson, C., Willson, G. A., . . . Andersen, M. E. (2008). Derivation of an inhalation reference concentration based upon olfactory neuronal loss in male rats following subchronic acetaldehyde inhalation. *Inhal Toxicol, 20*(3), 245-256. doi:10.1080/08958370701864250

Dourson, M. L., Teuschler, L. K., Durkin, P. R., & Stiteler, W. M. (1997). Categorical regression of toxicity data: a case study using aldicarb. *Regul Toxicol Pharmacol, 25*(2), 121-129. doi:10.1006/rtph.1996.1079

EFSA. (2010). Application of systematic review methodology to food and feed safety assessments to support decision mking. *EFSA Journal, 8*(6), 1637. Retrieved from https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2010.1637

EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal, 12*(6), 1831–4732. Retrieved from https://www.efsa.europa.eu/en/efsajournal/pub/3734

EFSA. (2015). Principles and process for dealing with data and evidence in scientific assessments. *EFSA Journal, 13*(5), 4121–4136. Retrieved from https://www.efsa.europa.eu/en/efsajournal/pub/4121

EPA. (1994). *Methods for Derivation of Inhalation Reference Concentrations (RfCs) and Application of Inhalation Dosimetry*. Washington, DC.: US EPA. Retrieved from https://www.epa.gov/risk/methods-derivation-inhalation-reference-concentrations-and-application-inhalation-dosimetry

EPA. (1995). *The Use of the Benchmark Dose Approach in Health Risk Assessment*. Washington, DC.: US EPA. Retrieved from http://hero.epa.gov/index.cfm/reference/download/reference_id/5992.

EPA. (2011). *Recommended Use of Body Weight 3/4 as the Default Method in Derivation of the Oral Reference Dose*. Washington, DC: US EPA. Retrieved from https://www.epa.gov/risk/recommended-use-body-weight-34-default-method-derivation-oral-reference-dose

EPA. (2012). Benchmark Dose Technical Guidance. Retrieved from https://www.epa.gov/sites/production/files/2015-01/documents/benchmark_dose_guidance.pdf

EPA. (2017). Categorical Regression (CATREG) User Guide Version 3.1.0.7. Retrieved from https://www.epa.gov/sites/production/files/2016-03/.../catreg_user_guide.pdf

EPA. (2018). Basic Information about the Integrated Risk Information System. Retrieved from https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system

EPA. (2019). List of Alternative Test Methods and Strategies (or New Approach Methodologies [NAMs]. Retrieved from https://www.epa.gov/sites/production/files/2019-12/documents/alternative_testing_nams_list_first_update_final.pdf

European Chemicals Agency. (2013). *Guidance for Human Health Risk Assessment* (Vol. 3, Part B). Retrieved from https://echa.europa.eu/documents/10162/23492134/biocides_guidance_vol_iii_part_b_v10_superseded_en.pdf/8ce06b02-2a0b-a348-7a44-162a8c83e633

European Food Safety Authority. (2015). Conclusion on the peer review of the pesticide risk assessment of the active substance glyphosate. *EFSA Journal, 13*(11), 4302. Retrieved from https://www.efsa.europa.eu/en/efsajournal/pub/4302

Farhat, N., Saunders-Hastings, P., Morgan, R., Tsaouin, K., Ramoju, S., & Krewski, D. (2020). Best Practices in Systematic Review. *Altex, under review*.

Farrell, P. J., Milton, B., Ramoju, S., Mattison, D., Birkett, N., & Krewski, D. (2020). The use of categorical regression in evidence integration. *Altex, under review*.

Farrugia, P., Petrisor, B. A., Farrokhyar, F., & Bhandari, M. (2010). Practical tips for surgical research: Research questions, hypotheses and objectives. *Canadian journal of surgery. Journal canadien de chirurgie, 53*(4), 278-281.

Fehringer, G., Brenner, D. R., Zhang, Z. F., Lee, Y. A., Matsuo, K., Ito, H., . . . Hung, R. J. (2017). Alcohol and lung cancer risk among never smokers: A pooled analysis from the international lung cancer consortium and the SYNERGY study. *Int J Cancer, 140*(9), 1976-1984. doi:10.1002/ijc.30618

Felix, A. S., Gaudet, M. M., La Vecchia, C., Nagle, C. M., Shu, X. O., Weiderpass, E., . . . Brinton, L. A. (2015). Intrauterine devices and endometrial cancer risk: a pooled analysis of the Epidemiology of Endometrial Cancer Consortium. *Int J Cancer, 136*(5), E410-422. doi:10.1002/ijc.29229

Field, R. W., Krewski, D., Lubin, J. H., Zielinski, J. M., Alavanja, M., Catalan, V. S., . . . Wilcox, H. B. (2006). An overview of the North American residential radon and lung cancer case-control studies. *J Toxicol Environ Health A, 69*(7), 599-631. doi:10.1080/15287390500260960

Friedenreich, C. M. (1993). Methods for pooled analyses of epidemiologic studies. *Epidemiology, 4*(4), 295-302. doi: 10.1097/00001648-199307000-00004

Gaudet, M. M., Olshan, A. F., Chuang, S. C., Berthiller, J., Zhang, Z. F., Lissowska, J., . . . Hashibe, M. (2010). Body mass index and risk of head and neck cancer in a pooled analysis of case-control studies in the International Head and Neck Cancer Epidemiology (INHANCE) Consortium. *Int J Epidemiol, 39*(4), 1091-1102. doi:10.1093/ije/dyp380

Gift, J. S., McGaughy, R., Singh, D. V., & Sonawane, B. (2008). Health assessment of phosgene: approaches for derivation of reference concentration. *Regul Toxicol Pharmacol, 51*(1), 98-107. doi:10.1016/j.yrtph.2008.03.004

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., . . . Schunemann, H. J. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol, 64*(4), 395-400. doi:10.1016/j.jclinepi.2010.09.012

Guyatt, G. H., Oxman, A. D., Kunz, R., Falck-Ytter, Y., Vist, G. E., Liberati, A., & Schunemann, H. J. (2008). Going from evidence to recommendations. *Bmj, 336*(7652), 1049-1051. doi:10.1136/bmj.39493.646875.AE

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., . . . Schunemann, H. J. (2011). GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol, 64*(12), 1303-1310. doi:10.1016/j.jclinepi.2011.04.014

Guyatt, G. H., Oxman, A. D., Sultan, S., Glasziou, P., Akl, E. A., Alonso-Coello, P., . . . Schunemann, H. J. (2011). GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol, 64*(12), 1311-1316. doi:10.1016/j.jclinepi.2011.06.004

Haber, L., Strickland, J., & Guth, D. J. (2001). *Categorical regression analysis of toxicity data* (Vol. 7). Retrieved from https://www.tera.org/Publications/catreg2001.pdf

Hattis, D., Baird, S., & Goble, R. (2002). A straw man proposal for a quantitative definition of the RfD. *Drug Chem Toxicol, 25*(4), 403-436. doi:10.1081/dct-120014793

Health Canada. (2016). Chemicals Management Plan Risk Assessment Toolbox. Retrieved from https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/chemicals-management-plan-risk-assessment-toolbox.html

Health Canada. (2018). Science Approach Documents. Retrieved from https://www.canada.ca/en/health-canada/services/chemical-substances/science-approach-documents.html.

HEI Diesel Epidemiology Panel. (2015). *Diesel Emissions and Lung Cancer: An Evaluation of Recent Epidemiological Evidence for Quantitative Risk Assessment*. Boston, MA: Health Effects Institute. Retrieved from https://www.healtheffects.org/publication/diesel-emissions-and-lung-cancer-evaluation-recent-epidemiological-evidence-quantitative

Hertzberg, R. C., & Miller, M. (1985). A statistical model for species extrapolation using categorical response data. *Toxicol Ind Health, 1*(4), 43-57. doi:10.1177/074823378500100405

Hooijmans, C. R., de Vries, R. B. M., Ritskes-Hoitinga, M., Rovers, M. M., Leeflang, M. M., IntHout, J., . . . Langendam, M. W. (2018). Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS ONE, 13*(1), e0187271. doi:10.1371/journal.pone.0187271

IARC. (1978b). IARC monographs on the evaluation of the carcinogenic risk of chemicals to humans: some N-nitroso compounds. *IARC Monogr Eval Carcinog Risk Chem Man, 17*, 1-349.

IARC. (2019). *Preamble to the IARC Monographs on the Identification of Carcinogenic Hazards to Humans*. Lyon: International Agency for Research on Cancer.

Johnson, P. I., Koustas, E., Vesterinen, H. M., Sutton, P., Atchley, D. S., Kim, A. N., . . . Woodruff, T. J. (2016). Application of the Navigation Guide systematic review methodology to the evidence for developmental and reproductive toxicity of triclosan. *Environ Int, 92-93*, 716-728. doi:10.1016/j.envint.2016.03.009

Johnson, P. I., Sutton, P., Atchley, D. S., Koustas, E., Lam, J., Sen, S., . . . Woodruff, T. J. (2014). The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth. *Environ Health Perspect, 122*(10), 1028-1039. doi:10.1289/ehp.1307893

Kadry Taher, M., Farhat, N., Karyakina, N. A., Shilnikova, N., Ramoju, S., Gravel, C. A., . . . Krewski, D. (2019). Critical review of the association between perineal use of talc powder and risk of ovarian cancer. *Reprod Toxicol, 90*, 88-101. doi:10.1016/j.reprotox.2019.08.015

Kheifets, L., Ahlbom, A., Crespi, C. M., Draper, G., Hagihara, J., Lowenthal, R. M., . . . Wunsch Filho, V. (2010). Pooled analysis of recent studies on magnetic fields and childhood leukaemia. *Br J Cancer, 103*(7), 1128-1135. doi:10.1038/sj.bjc.6605838

Koustas, E., Lam, J., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., . . . Woodruff, T. J. (2014). The Navigation Guide - evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environ Health Perspect, 122*(10), 1015-1027. doi:10.1289/ehp.1307177

Krewski, D., Anderson, M., Tyshenko, M., Krishnan, K., Hartung, T., Boekelheide, K., . . . Cote, I. (2019). Toxicity Testing in the 21st Century: Progress in the Past Decade and Future Perspectives. *Archives of Toxicology, in press*.

Krewski, D., Barakat-Haddad, C., Donnan, J., Martino, R., Pringsheim, T., Tremlett, H., . . . Cashman, N. R. (2017). Determinants of neurological disease: Synthesis of systematic reviews. *Neurotoxicology, 61*, 266-289. doi:10.1016/j.neuro.2017.04.002

Krewski, D., Chambers, A., Stern, B. R., Aggett, P. J., Plunkett, L., & Rudenko, L. (2010). Development of a copper database for exposure-response analysis. *J Toxicol Environ Health A, 73*(2), 208-216. doi:10.1080/15287390903340815

Krewski, D., Lubin, J. H., Zielinski, J. M., Alavanja, M., Catalan, V. S., Field, R. W., . . . Wilcox, H. B. (2006). A combined analysis of North American case-control studies of residential radon and lung cancer. *J Toxicol Environ Health A, 69*(7), 533-597. doi:10.1080/15287390500260945

Krewski, D., Westphal, M., Andersen, M. E., Paoli, G. M., Chiu, W. A., Al-Zoughool, M., . . . Cote, I. (2014). A framework for the next generation of risk science. *Environ Health Perspect, 122*(8), 796-805. doi:10.1289/ehp.1307260

Lam, J., Koustas, E., Sutton, P., Johnson, P. I., Atchley, D. S., Sen, S., . . . Woodruff, T. J. (2014). The Navigation Guide - evidence-based medicine meets environmental health: integration of animal and human evidence for PFOA effects on fetal growth. *Environ Health Perspect, 122*(10), 1040-1051. doi:10.1289/ehp.1307923

Lam, J., Lanphear, B. P., Bellinger, D., Axelrad, D. A., McPartland, J., Sutton, P., . . . Woodruff, T. J. (2017). Developmental PBDE Exposure and IQ/ADHD in Childhood: A Systematic Review and Meta-analysis. *Environ Health Perspect, 125*(8), 086001. doi:10.1289/ehp1632

Lam, J., Sutton, P., Kalkbrenner, A., Windham, G., Halladay, A., Koustas, E., . . . Woodruff, T. (2016). A Systematic Review and Meta-Analysis of Multiple Airborne Pollutants and Autism Spectrum Disorder. *PLoS ONE, 11*(9), e0161851. doi:10.1371/journal.pone.0161851

Lehman, A. J. (1954). 100-fold margin of safety. *Quarterly Bulletin of Food and Drug Officials, 18*, 33-35. Retrieved from https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/3195

Lucas, R. M., & McMichael, A. J. (2005). Association or causation: evaluating links between "environment and disease". *Bull World Health Organ, 83*(10), 792-795. doi:s0042-96862005001000017

Makowski, D., Albert, I., Bonvallot, N., Boudia, S., Brochot, C., & et al. (2016). Opinion of the French Agency for Food, Environmental and Occupational Health & Safety regarding the progress report on the assessment of the weight of evidence at ANSES: critical literature review and recommendations at the hazard identification stage. *ANSES Opinion Request, No 2015-SA-0089*.

Martin, P., Bladdier, C., Meek, B., Bruyere, O., Feinblatt, E., Touvier, M., . . . Makowski, D. (2018). Weight of evidence for hazard identification: a critical review of the literature. *Environmental Health Perspectives, 126*(7), 1–15. doi: 10.1289/EHP3067

Masic, I., Miokovic, M., & Muhamedagic, B. (2008). Evidence based medicine - new approaches and challenges. *Acta Inform Med, 16*(4), 219-225. doi:10.5455/aim.2008.16.219-225

Milton, B., Farrell, P. J., Birkett, N., & Krewski, D. (2017a). Modeling U-Shaped Exposure-Response Relationships for Agents that Demonstrate Toxicity Due to Both Excess and Deficiency. *Risk Anal, 37*(2), 265-279. doi:10.1111/risa.12603

Milton, B., Krewski, D., Mattison, D. R., Karyakina, N. A., Ramoju, S., Shilnikova, N., . . . McGough, D. (2017b). Modeling U-shaped dose-response curves for manganese using categorical regression. *Neurotoxicology, 58*, 217-225. doi:10.1016/j.neuro.2016.10.001

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med, 6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Morgan, R. L., Thayer, K. A., Santesso, N., Holloway, A. C., Blain, R., Eftim, S. E., . . . Schunemann, H. J. (2019). A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. *Environ Int, 122*, 168-184. doi:10.1016/j.envint.2018.11.004

Morgan, R. L., Whaley, P., Thayer, K. A., & Schunemann, H. J. (2018). Identifying the PECO: A framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. *Environ Int, 121*(Pt 1), 1027-1031. doi:10.1016/j.envint.2018.07.015

Munafo, M., Nosek, B., Bishop, D., Button, K., Chambers, C., du Sert, N., . . . Ionnidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*. doi:10.1038/s41562-016-0021

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evidence Based Medicine, 21*(4), 125. doi:10.1136/ebmed-2016-110401

National Academies of Science. (2001). *Standard Operating Procedures for Developing Acute Exposure Guideline Levels*. Washington, DC.: National Academies Press. Retrieved from https://www.nap.edu/catalog/10122/standing-operating-procedures-for-developing-acute-exposure-guideline-levels-for-hazardous-chemicals

National Academies of Sciences Engineering and Medicine. (2018). *Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation*. Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/catalog/25086/progress-toward-transforming-the-integrated-risk-information-system-iris-program

NRC Committee on Improving Risk Analysis Approaches Used by the US EPA. (2009). *Science and Decisions: Advancing Risk Assessment*. Washington (DC): National Academies Press. Retrieved from https://www.nap.edu/catalog/12209/science-and-decisions-advancing-risk-assessment

NRC Committee on Risk Assessment of Hazardous Air Pollutants. (1994). *Science and Judgement in Risk Assessment*. Washington, DC.: National Academies Press. Retrieved from https://www.nap.edu/catalog/2125/science-and-judgment-in-risk-assessment

NRC. (2007). Toxicity Testing in the 21st Century: A Vision and Strategy. Retrieved from https://www.nap.edu/catalog/11970/toxicity-testing-in-the-21st-century-a-vision-and-a

NRC. (2008). *Phthalates and Cumulative Risk Assessment: The Task Ahead*. Washington, DC: National Academies Press. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25009926

NRC. (2011). *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde*. Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/catalog/13142/review-of-the-environmental-protection-agencys-draft-iris-assessment-of-formaldehyde

NRC. (2014a). *Review of EPA's Integrated Risk Information System (IRIS) Process*. Washington, DC: National Academics Press. Retrieved from https://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process

NRC. (2014b). *Review of the Environmental Protection Agency's State-of-the-science Evaluation of Nonmonotonic Dose-response Relationships as They Apply to Endocrine Disruptors*. Washington, DC: National Academies Press. Retrieved from https://www.nap.edu/catalog/18608/review-of-the-environmental-protection-agencys-state-of-the-science-evaluation-of-nonmonotonic-dose-response-relationships-as-they-apply-to-endocrine-disruptors

OECD. (2019). Case study on the use of an integrated approach to testing and assessment for estrogen receptor active chemicals. Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2019)28&docLanguage=en

OECD. (2020). OECD QSAR Toolbox. Retrieved from https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm

Peres, L. C., Risch, H., Terry, K. L., P, M. W., Goodman, M. T., Wu, A. H., . . . Schildkraut, J. M. (2018). Racial/ethnic differences in the epidemiology of ovarian cancer: a pooled analysis of 12 case-control studies. *Int J Epidemiol*. doi:10.1093/ije/dyy054

Petrisor, B., & Bhandari, M. (2007). The hierarchy of evidence: Levels and grades of recommendation. *Indian Journal of Orthopaedics, 41*(1), 11-15. doi:10.4103/0019-5413.30519

Rhomberg, L. R., Goodman, J. E., Bailey, L. A., Prueitt, R. L., Beck, N. B., Bevan, C., . . . Becker, R. A. (2013). A survey of frameworks for best practices in weight-of-evidence analyses. *Crit Rev Toxicol, 43*(9), 753-784. doi:10.3109/10408444.2013.832727

Rooney, A. A., Boyles, A. L., Wolfe, M. S., Bucher, J. R., & Thayer, K. A. (2014). Systematic review and evidence integration for literature-based environmental health science assessments. *Environ Health Perspect, 122*(7), 711-718. doi:10.1289/ehp.1307972

Sackett, D. L. (1997). Evidence-based medicine. *Semin Perinatol, 21*(1), 3-5. doi: 10.1016/s0146-0005(97)80013-4

Samet, J. M., Chiu, W. A., Cogliano, V., Jinot, J., Kriebel, D., Lunn, R. M., . . . Wild, C. P. (2019). The IARC Monographs: Updated procedures for modern and transparent evidence synthesis in cancer hazard identification. *J Natl Cancer Inst*. doi:10.1093/jnci/djz169

Sand, S. (2020). A Novel Method for Combining Outcomes with Different Severities. *Altex, under review*.

Sand, S., Lindqvist, R., von Rosen, D., & Ilback, N. G. (2018). Dose-Related Severity Sequence, and Risk-Based Integration, of Chemically Induced Health Effects. *Toxicol Sci, 165*(1), 74-89. doi:10.1093/toxsci/kfy124

Sand, S., Portier, C. J., & Krewski, D. (2011). A signal-to-noise crossover dose as the point of departure for health risk assessment. *Environ Health Perspect, 119*(12), 1766-1774. doi:10.1289/ehp.1003327

Saunders-Hastings, P., Rhomberg, L., & Krewski, D. (2020). Best Practices in Weight of Evidence Analysis: A Review. *Altex, under review*.

Schroeter, J. D., Kimbell, J. S., Gross, E. A., Willson, G. A., Dorman, D. C., Tan, Y. M., & Clewell, H. J., 3rd. (2008). Application of physiological computational fluid dynamics models to predict interspecies nasal dosimetry of inhaled acrolein. *Inhal Toxicol, 20*(3), 227-243. doi:10.1080/08958370701864235

Schunemann, H. J., Cuello, C., Akl, E. A., Mustafa, R. A., Meerpohl, J. J., Thayer, K., . . . Guyatt, G. (2019). GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol, 111*, 105-114. doi:10.1016/j.jclinepi.2018.01.012

Schunemann, H. J., Oxman, A. D., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G. E., . . . Guyatt, G. H. (2008). Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Bmj, 336*(7653), 1106-1110. doi:10.1136/bmj.39500.677199.AE

Schünemann, H. J., Zhang, Y., & Oxman, A. D. (2019). Distinguishing opinion from evidence in guidelines. *Bmj, 366*, l4606. doi:10.1136/bmj.l4606

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *Bmj, 350*. doi: https://doi.org/10.1136/bmj.g7647

Smith, M. T., Guyton, K. Z., Gibbons, C. F., Fritz, J. M., Portier, C. J., Rusyn, I., . . . Straif, K. (2016). Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ Health Perspect, 124*(6), 713-721. doi:10.1289/ehp.1509912

Stephens, M. L., Andersen, M., Becker, R. A., Betts, K., Boekelheide, K., Carney, E., . . . Zurlo, J. (2013). Evidence-based toxicology for the 21st century: opportunities and challenges. *ALTEX, 30*(1), 74-103. doi:10.14573/altex.2013.1.074

Stern, B. R., Solioz, M., Krewski, D., Aggett, P., Aw, T. C., Baker, S., . . . Starr, T. (2007). Copper and human health: biochemistry, genetics, and strategies for modeling dose-response relationships. *J Toxicol Environ Health B Crit Rev, 10*(3), 157-222. doi:10.1080/10937400600755911

Teeguarden, J. G., Bogdanffy, M. S., Covington, T. R., Tan, C., & Jarabek, A. M. (2008). A PBPK model for evaluating the impact of aldehyde dehydrogenase polymorphisms on comparative rat and human nasal tissue acetaldehyde dosimetry. *Inhal Toxicol, 20*(4), 375-390. doi:10.1080/08958370801903750

The National Academies of Sciences Engineering and Medicine. (2017). *Guiding Principles for Developing Dietary Reference Intakes Based on Chronic Diseases*. Washington, DC: NASEM. Retrieved from http://www.nationalacademies.org/hmd/Reports/2017/guiding-principles-for-developing-dietary-reference-intakes-based-on-chronic-disease.aspx

Thomas, D. C., Jerrett, M., Kuenzli, N., Louis, T. A., Dominici, F., Zeger, S., . . . Bates, D. (2007). Bayesian model averaging in time-series studies of air pollution and mortality. *J Toxicol Environ Health A, 70*(3-4), 311-315. doi:10.1080/15287390600884941

Thomas, R. S., Bahadori, T., Buckley, T. J., Cowden, J., Deisenroth, C., Dionisio, K. L., . . . Williams, A. J. (2019). The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. *Toxicological Sciences, 169*(2), 317-332. doi:10.1093/toxsci/kfz058

Thomas, R. S., Philbert, M. A., Auerbach, S. S., Wetmore, B. A., Devito, M. J., Cote, I., . . . Nong, A. (2013). Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. *Toxicol Sci, 136*(1), 4-18. doi:10.1093/toxsci/kft178

Tobias, A., Saez, M., & Kogevinas, M. (2004). Meta-analysis of results and individual patient data in epidemiological studies. *Journal of Modern Applied Statistical Methods, 3*(1), 176–185. Retrieved from https://pdfs.semanticscholar.org/819f/aa4d45cbdacc501d97293e1a4914c3a175c1.pdf

Vermeulen, R., Silverman, D. T., Garshick, E., Vlaanderen, J., Portengen, L., & Steenland, K. (2014). Exposure-response estimates for diesel engine exhaust and lung cancer mortality based on data from three occupational cohorts. *Environ Health Perspect, 122*(2), 172-177. doi:10.1289/ehp.1306880

Vesterinen, H. M., Johnson, P. I., Atchley, D. S., Sutton, P., Lam, J., Zlatnik, M. G., . . . Woodruff, T. J. (2015). Fetal growth and maternal glomerular filtration rate: a systematic review. *J Matern Fetal Neonatal Med, 28*(18), 2176-2181. doi:10.3109/14767058.2014.980809

Wang, M. D., Gomes, J., Cashman, N. R., Little, J., & Krewski, D. (2014). A meta-analysis of observational studies of the association between chronic occupational exposure to lead and amyotrophic lateral sclerosis. *J Occup Environ Med, 56*(12), 1235-1242. doi:10.1097/jom.0000000000000323

Webster, F., Gagné, M., Patlewicz, G., Pradeep, P., Trefiak, N., Judson, R. S., & Barton-Maclaren, T. S. (2019). Predicting estrogen receptor activation by a group of substituted phenols: An integrated approach to testing and assessment case study. *Regulatory Toxicology and Pharmacology, 106*, 278-291. doi:https://doi.org/10.1016/j.yrtph.2019.05.017

West, G. B. (1999). The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science, 284*(5420), 1677-1679. doi:10.1126/science.284.5420.1677

WHO/IPCS. (2009). *Principles for modelling dose-response for the risk assessment of chemicals*. Geneva, Switzerland: World Health Organization International Program on Chemical Safety. Retrieved from https://apps.who.int/iris/handle/10665/43940

WHO/IPCS. (2014). *Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization*. Geneva, Switzerland: World Health Organization International Program on Chemical Safety. Retrieved from https://apps.who.int/iris/handle/10665/259858

Wigle, D. T., Arbuckle, T. E., Turner, M. C., Berube, A., Yang, Q., Liu, S., & Krewski, D. (2008). Epidemiologic evidence of relationships between reproductive and child health outcomes and

environmental chemical contaminants. *J Toxicol Environ Health B Crit Rev, 11*(5-6), 373-517. doi:10.1080/10937400801921320

Woodruff, T. J., & Sutton, P. (2011). An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. *Health Aff (Millwood), 30*(5), 931-937. doi:10.1377/hlthaff.2010.1219

Wyss, A., Hashibe, M., Chuang, S. C., Lee, Y. C., Zhang, Z. F., Yu, G. P., . . . Olshan, A. F. (2013). Cigarette, cigar, and pipe smoking and the risk of head and neck cancers: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Am J Epidemiol, 178*(5), 679-690. doi:10.1093/aje/kwt029

Yetley, E. A., MacFarlane, A. J., Greene-Finestone, L. S., Garza, C., Ard, J. D., Atkinson, S. A., . . . Wells, G. A. (2017). Options for basing Dietary Reference Intakes (DRIs) on chronic disease endpoints: report from a joint US-/Canadian-sponsored working group. *Am J Clin Nutr, 105*(1), 249s-285s. doi:10.3945/ajcn.116.139097

Supplementary Material
Development of an Evidence Based Risk Assessment Framework
Workshop Agenda

Monday, December 17th, 2018

Welcome and Overview

8:15 am – 8:20 am	Guy Levesque, Associate Vice President, University of Ottawa Welcome Welcome
8:20 am – 8:30 am	Daniel Krewski, University of Ottawa Risk science in the 21st century: Overview

Session 1: Recent Advances in Risk Science: Including New Approach Methodologies in Weight of Evidence Evaluation

This session will take stock of recent scientific developments that will support evidence-based risk assessment, including new approach methodologies (NAMs).

Chair: Thomas Hartung, Johns Hopkins University

8:30–8:55 am	Maureen Gwinn, US EPA Current Status of New Approach Methodologies
8:55 am – 9:20 am	Patience Browne. OECD Predictive value of in vitro assays
9:20 am – 9:45 am	Andrew Rooney, NIEHS Incorporating information from new approach methodologies in weight of evidence evaluation)
9:45 am – 10: am	General discussion
10:00 am – 10:30 am	*Break*

Session 2: Summarizing the Evidence

This session will focus on methods for summarizing all relevant data to be included in an evidence-based risk assessment. Methods in systematic review will be examined, along with current approaches to data quality scoring.

Chair: Jeff Lewis, Exxon Mobil Biomedical Research

10:30 am – 10:55 am	Juleen Lam, Cal State East Bay Integrating multiple evidence streams

10:55 am – 11:20 am	Thomas Hartung, Johns Hopkins Systematic review of toxicological data
11:20 am – 11:45 am	Charlotte Bertrand, US EPA Quality scoring of human, animal, and in vitro data
11:45 am – 12:00 pm	General discussion
12:00 pm – 1:00 pm	*Lunch* Demonstration of Bayesian Weight of Evidence Decision-Support Tool Moez Sanaa, ANSES and Greg Paoli, Risk Sciences International

Session 3: Qualitative Data Synthesis

The first step in evidence-based risk assessment is the determination of whether or not a hazard exists. This involves a weight of evidence evaluation of all relevant information in order to reach a decision on whether the available data supports the existence of a human health hazard.

Chair: Kristina Thayer, US EPA

1:00 pm – 1:25 pm	Kurt Straif (confirmed, IARC The IARC Monographs Programme of identification of carcinogenic hazards to humans
1:25 pm – 1:50 pm	Holger Schünemann, McMaster Use of GRADE in evidence integration
1:50 pm – 2:15 pm	Andrew Kraft, US EPA Current and future EPA practices in systematic review
2:15 pm – 2:30 pm	General discussion
2:30 pm – 3:00	*Break*

Session 4: Quantitative Data Synthesis

Once a hazard has been identified on the basis of the available evidence, a quantitative assessment of risk and exposure-response may be undertaken. This session will focus on new methodologies for quantitative synthesis of data from multiple sources, including synthesis of data on diverse toxicological endpoints.

Chair: Greg Paoli

3:00 pm – 3:25 pm	Salomon Sand, Swedish National Food Agency New approaches for quantitative combining of data from multiple sources
3:25 pm – 3:50 pm	Don Mattison, Risk Sciences International Quantitative synthesis of neurotoxicity data on manganese using categorical regression

3:50 pm – 4:15 pm	Weihsueh Chiu, Texas A&M University New approaches to characterizing uncertainty in risk assessment
4:15 pm – 4:40 pm	Katya Tsaioun, Johns Hopkins University In vitro predictions of drug induced liver injury
4:40 pm – 5:00 pm	General discussion
5:00 pm	*Adjourn*

Tuesday, December 18th, 2018

8:30 am – 9:00 am	Summary of Day 1 Daniel Krewski, University of Ottawa

Session 5: Putting Weight of Evidence into Practice

In order to guide discussions about considerations involved in the practical implementation of weight of evidence, this session will provide an overview of current approaches within EFSA and Health Canada.
Chair: Maureen Gwinn, EPA

9:00 am – 9:25 am	Elisa Aiassa, Laura Martino and Caroline Merten, EFSA Evidence integration: an EU perspective
9:25 am – 9:50 am	Tara-Barton Maclaren, Health Canada Health Canada's evolving framework for evidence synthesis
10:00 am – 10:30 am	*Break*

The remainder of the meeting will be held in closed session.

Session 6: Breakout Groups

Participants at the workshop will be assigned to breakout groups to address a series of key questions relating to the development of an evidence-based framework for risk assessment. (Questions developed by the Steering Committee.)
Moderator: Tara Barton-Maclaren, Health Canada

10:30 am – 12:00 pm	Parallel Breakout Group Discussions
	Group 1: Lessons learned from previous experience Chair: Lorenz Rhomberg, Gradient Corporation Rapporteur: Patrick Saunders-Hastings, Gevity

Group 2: Benchmarks of good practice
Chair: Greg Paoli, Risk Sciences International
Rapporteur: Maureen Gwinn, US EPA

Group 3: Problem formulation and data requirements
Chair: Robert Baan, IARC (retired)
Rapporteur: Kris Thayer, US EPA

Group 4: Potential challenges
Chair: Thomas Hartung, Johns Hopkins
Rapporteur: Rebecca Morgan, McMaster University

12:00 pm – 1:00 pm *Lunch*

1:00 pm – 2:00 pm Breakout Group Reports
Moderator: Tara Barton-Maclaren, Health Canada

2:00 pm – 2:30 pm Synthesis of Breakout Group Reports
Daniel Krewski, University of Ottawa

2:30 pm – 3:00 pm *Break*

Session 6: General Discussion and Next Steps

This session will include a general discussion of key themes identified at the workshop and possible components of an evidence-based risk assessment framework. (Steering Committee members will be asked to provide their perspectives on future directions, with input from participants.)
Chair: Thomas Hartung, Johns Hopkins

3:00 pm – 3:30 pm Opening 5-minute presentations by Steering Committee members:
Tara Barton-Maclaren, Health Canada; Thomas Hartung, Johns Hopkins University; Daniel Krewski, University of Ottawa; Kristina Thayer, US EPA; Jeff Lewis, Exxon Mobil Biomedical Research.

3:30 pm – 4:00 pm General discussion

4:00 pm – 4:30 pm Conclusion
Daniel Krewski, University of Ottawa

4:30 pm *Adjourn*